Striatal cell-type specific encoding stability and reorganization for agency and habit

Melissa Malvaez[1], Alvina Liang[1], Baila S. Hall[1], Jacqueline R. Giovanniello[1], Natalie Paredes[1], Julia Y. Gonzalez[1], Garrett J. Blair[1], Ana C. Sias[1], Michael D. Murphy[1], Wanyi Guo[1], Alicia Wang[1], Malika Singh[1], Nicholas K. Griffin[1], Samuel P. Bridges[1], Anna Wiener[1], Jenna S. Pimenta[1] Sandra M. Holley[3], Carlos Cepeda[2, 3], Michael S. Levine[2, 3], Peter D. Balsam[4,5], H. Tad Blair[1, 2], Andrew M. Wikenheiser[1, 2], Kate M. Wassum[1, 2]

[1]Dept. of Psychology, UCLA, Los Angeles, CA 90095. [2]Brain Research Institute, UCLA, Los Angeles, CA 90095, USA.[3]Intellectual and Developmental Disabilities Research Center, Semel Institute for Neuroscience and Human Behavior, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, United States, [4]Department of Psychology, Barnard College, New York City, NY, USA. [5]Columbia University and New York State Psychiatric Institute, New York City, NY, USA.

Correspondence:
Kate Wassum: kwassum@ucla.edu
Dept. of Psychology, UCLA
1285 Franz Hall
Box 951563
Los Angeles, CA 90095-1563

Key words: learning, decision making, instrumental conditioning, reward, striatum, habit, devaluation, calcium imaging, chemogenetics, direct and indirect pathway

Figures: 6
Tables: 0
Extended Data Figures: 32
Supplemental Tables: 4

SUMMARY

Adaptive, goal-directed decision making requires agency, knowledge that an action will cause a particular outcome. For well-practiced routines, agency is relinquished in favor of automatic habits. Here, we used cell-type specific, longitudinal, cellular-resolution, microendoscopic calcium imaging, chemogenetics, instrumental conditioning and a test of behavioral strategy to ask how dorsomedial striatum $D1^+$ and $D2/A2A^+$ neurons contribute to the formation of agency and habit. We found that DMS $D1^+$ neurons stably encode actions and variables related to goal-directed v. habitual behavioral strategy, develop encoding of the action-outcome relationship, and are critical for both action-outcome learning and the application of this learned agency for flexible, goal-directed decision making. Conversely, one ensemble of $A2A^+$ neurons transiently encodes actions and behavioral transition variables during early action-outcome learning and another slowly starts to encode actions as routine habits form. Correspondingly, $A2A^+$ neurons are only transiently necessary for initial formation of the action-outcome model underlying agency. Thus, agency is learned through DMS $D1^+$ and $A2A^+$ neuronal activity and then maintained in DMS $D1^+$ neurons to support adaptive, goal-directed decision making. Habit formation is associated with reorganization of $A2A^+$ neuronal encoding. These data reveal important insights into the neuronal circuits of learning and decision making.

When making a decision, we often engage our agency. Our brains use knowledge of available actions and their consequences to prospectively evaluate options and choose actions that will cause beneficial outcomes[1-3]. This goal-directed strategy is model based. We build an internal model of action-outcome relationships to support the predictions and inferences needed for adaptive decision making[2-8]. It is, therefore, highly flexible. Should an outcome become unbeneficial, we will reduce the actions that cause it. But it is also cognitively taxing. So, we have a model-free strategy for well-practiced routines: habits. Habits typically form slowly. With repeated successful practice, behavioral strategy transitions from reliance on action-outcome knowledge to more automatic execution[2,9-12]. Habits are, thus, resource efficient, but inflexible. We might continue a habit even when it does not serve our goals. Balance between goal-directed and habitual control allows behavior to be adaptive when needed, yet efficient when appropriate[13,14]. Disrupted agency and overreliance on habit can, however, cause inadequate consideration of consequences, disrupted decision making, inflexible behavior, and a lower threshold for compulsivity[15-18]. This can contribute to cognitive deficits in addiction[19-30] and myriad other mental health conditions[17,26,31-40]. Yet, despite importance for understanding adaptive and maladaptive behavior, much is unknown of how the brain forms action-outcome associations to support agency and what changes as habits form.

The dorsomedial striatum (DMS) is an evolutionally conserved hub for action-outcome learning and goal-directed decision making[9,41-52]. In the striatum, experiential factors that drive learning converge onto basal ganglia outputs that drive motor execution[53]. There are two major subtypes of these outputs: one expressing dopamine D1 receptors and another expressing dopamine D2 and adenosine Adora2A (A2A) receptors. Whereas D1$^+$ neurons are more dominant in the direct pathway to basal ganglia output nuclei, D2/A2A$^+$ neurons are associated with the multisynaptic, indirect output pathway[54-56]. Though this organization is likely more complex than originally thought[57,58]. The activity of these neurons has been well characterized during locomotion and previously learned actions[9,41,59]. But little is known of how their activity is shaped by action-outcome learning or transitions in behavioral strategy. Such information is needed to understand how learning and repeated practice of an action shape striatal activity to support the internal action-outcome model and enable the execution of behavioral strategies. Here we fill this gap in knowledge by tracking and manipulating the activity of DMS D1$^+$ and A2A$^+$ neurons during action-outcome learning and habit formation.

RESULTS

The DMS D1$^+$ neuronal population encodes behavior, information underlying behavioral strategy, and the action-outcome association.

We first asked how DMS D1$^+$ neuronal activity relates to action-outcome learning and the transition to habit. We used UCLA miniscopes[60] with a gradient refractive index (GRIN) lens for cellular-resolution, microendoscopic imaging of the genetically encoded calcium indicator jGCaMP7s[61] selectively expressed in D1$^+$ neurons in the DMS of Drd1a-cre mice (Figure 1a-d). We recorded calcium activity as mice learned to lever press to earn food-pellet rewards in a self-initiated instrumental task (Figure 1e). We used a random-interval reinforcement schedule that enables action-outcome learning for flexible goal-directed decision making with 4 training sessions and promotes the formation of inflexible habits with overtraining (8 sessions; Extended Data Figure 1-1). All mice acquired the

instrumental behavior (Figure 1f; see Extended Data Figure 1-2 for food-port entry data). To evaluate behavioral strategy, we used the outcome-specific devaluation test[2,62]. Mice were given 90-min, non-contingent access to the food pellet earned during training to induce a sensory-specific satiety rendering that pellet temporarily devalued. Lever pressing was then assessed in an immediately following 5-min, non-reinforced probe test. Performance was compared to that following satiation on an alternate food pellet to control for general satiety (Valued state). Goal-directed decisions involve the use of agency, knowledge of the action-outcome relationship[1-3]. This strategy dominates after 4 training sessions, such that mice reduce lever-press actions when their outcome is devalued (Extended Data Figure 1-1). Habits are marked by lack of consideration of action consequences, resulting in inflexibility to devaluation[2,9-12]. As expected, following overtraining, mice were insensitive to devaluation (Figure 1g-h; see also Extended Data Figure 1-1). To capture activity related to action-outcome learning, we analyzed calcium activity at the beginning (random-interval training session 1) and middle (session 4) of instrumental training. To evaluate activity related to habit, we analyzed calcium activity at the end of training (overtraining, session 8; see Supplemental Table 1 for neurons/mouse/session). We leveraged the ability to track neurons longitudinally to ask how DMS D1$^+$ neurons change their activity and encoding with action-outcome learning and habit formation (Figure 1d; Extended Data Figure 1-3; Supplemental Table 1).

Instrumental actions, reward outcomes, and information underlying behavioral strategy can be decoded from DMS D1$^+$ neuron population activity. We first used decoding to probe the information content available in DMS D1$^+$ neuronal population activity. We trained a multi-class support vector machine (SVM) to classify behavioral events from the temporal fluorescence traces of all coregistered D1$^+$ neurons. Behavioral events were *action initiation*: first lever press in the session, after earned reward collection, or after a non-reinforced food-port check; *action termination*: last press before reward collection or non-reinforced food-port check; *check*: non-reinforced food-port entry; and *reward:* earned reward consumption. Across training, the decoder classified each behavior more accurately than one trained on shuffled data (Figure 1i). Decoding of action initiation improved during initial training, but did not improve further with overtraining. Thus, the D1$^+$ neuronal population can convey information about actions and outcomes.

We next asked the extent to which information related to behavioral strategy could be decoded from DMS D1$^+$ neuronal population activity. To identify these variables, we first built a linear regression model to ask whether and how behavioral variables during the final training session predicted subsequent sensitivity to outcome devaluation (Extended Data Figure 1-4). We identified 4 important variables, 2 behavioral engagement variables: press rate (lever presses/min) and switch rate (rate at which subject switch from pressing to checking the food port or vice versa) and 2 behavioral organization variables: overall interpress interval variability/consistency (standard deviation of the log-transformed interpress intervals) and the extent to which current trial behavior relates to immediate past trial behavior (correlation between the current interpress interval, n, and the interpress interval of the prior press, n-1[63,64]). Together these variables predicted sensitivity to devaluation with very high accuracy (Extended Data Figure 1-4). Goal-directed actions were associated with being on task, engaged in pressing and checking for the reward and re-engagement in pressing (high press and switch rate) that is planned based on, or at least related to, past performance (low interpress interval variability and high n:n-1 interpress

interval correlation). Habits, on the other hand, were associated with less task engagement (lower press and switch rate) or less flexible task engagement (high press rate x low switch rate) and less consideration of past successful behavior (high interpress interval variability and low n:n-1 interpress interval correlation). We next trained linear regression models to decode these behavioral strategy variables from $D1^+$ neuronal population activity (see Extended data Figure 1-5). Press rate, switch rate, log interpress interval, and past trial interpress interval could all be decoded with high accuracy across training (Figure 1j-m). Thus, at the population level, DMS $D1^+$ neurons convey behavioral engagement and organizational information related to the extent of goal-directed v. habitual behavioral control.

The DMS $D1^+$ neuronal population represents the action-outcome relationship with learning. We next asked whether the DMS $D1^+$ neuronal population encodes the action-outcome relationship that underlies agency for goal-directed decision making. If so, we reasoned the $D1^+$ neuronal population activity trajectory of the earned reward should become more similar to that of the action with training. We used a population vector analysis[65] to examine the geometric relationship between the $D1^+$ neuronal population representations of action initiation and earned reward outcome during each phase of training (Figure 1n-q; see also Extended Data Figure 1-6). We defined the activity of each neuron as a dimension and the population response to each event as a vector in high-dimensional state space and used the cosine distance between the two vectors as a measure of the similarity in the pattern of population activity, with little sensitivity to the overall firing rate or vector magnitude (Figure 1n). The distance between the action initiation and reward population vectors decreased with training (Figure 1o-p). Thus, with learning, the DMS $D1^+$ neuronal population vectors representing action and earned reward outcome move closer together in representational space, indicating emergent action-outcome encoding. We found similar results using a representational similarity analysis (Extended Data Figure 1-7).

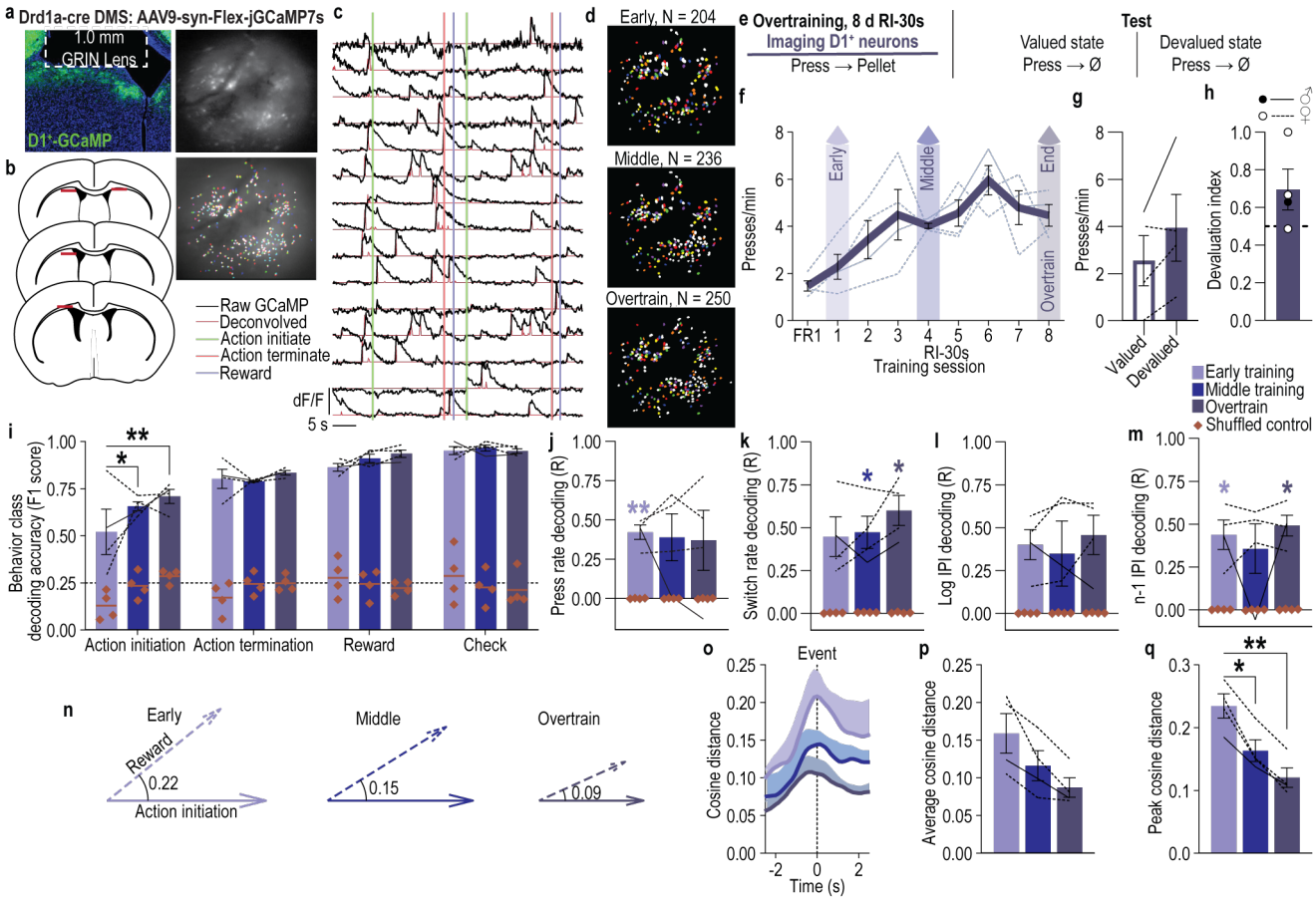


Figure 1: The DMS D1⁺ neuronal population encodes instrumental actions, reward outcomes, information underlying behavioral strategy, and the action-outcome association. (a) Representative cre-dependent jGCaMP7s expression in DMS D1⁺ neurons (top left) and maximum intensity projection of the field-of-view with (top right) and without (bottom right) demarcation of cell ROIs for one session. **(b)** Map of DMS GRIN lens placements. **(c)** Representative dF/F and deconvolved calcium signal. **(d)** Representative recorded D1⁺ neuron spatial footprints during the first (early), 4th (middle), and 8th (overtrain) random-interval training sessions. Colored, co-registered neurons; white, non-co-registered neurons. **(e)** Procedure. RI, random-interval reinforcement schedule. **(f)** Training press rate. 1-way ANOVA, Training: $F_{1.81, 5.43} = 4.84$, $P = 0.06$. **(g)** Test press rate. 2-tailed t-test, $t_3 = 1.98$, $P = 0.14$, 95% confidence interval (CI) -0.85 - 3.65. **(h)** Devaluation index [(Devalued presses)/(Valued presses + Devalued presses)]. One-tailed Bayes factor, $BF_{10} = 0.18$. **(i)** Behavior class decoding accuracy from D1⁺ coregistered neuron activity in 0.5-s period prior to event compared to shuffled control. Line at 0.25 = chance. 3-way ANOVA, Neuron activity (v. shuffled): $F_{1, 3} = 1092.83$, $P < 0.001$; Training session: $F_{1.17, 3.51} = 2.40$, $P = 0.21$; Behavior class: $F_{1.76, 5.28} = 15.56$, $P = 0.007$; Neuron activity x Training: $F_{1.24, 3.72} = 4.60$, $P = 0.10$; Neuron activity x Behavior: $F_{1.81, 5.43} = 6.49$, $P = 0.04$; Training x Behavior: $F_{2.15, 6.46} = 3.25$, $P = 0.10$; Neuron activity x Training x Behavior: $F_{1.95, 5.86} = 0.69$, $P = 0.54$. **(j-m)** Linear decoding accuracy from D1⁺ coregistered neuron activity of press rate (j; 2-way ANOVA, Neuron activity: $F_{1, 3} = 11.46$, $P = 0.04$; Training: $F_{1.22, 3.66} = 0.07$, $P = 0.85$; Neuron activity x Training: $F_{1.22, 3.66} = 0.07$, $P = 0.85$), switch rate (k; 2-way ANOVA, Neuron activity: $F_{1, 3} = 36.96$, $P = 0.009$; Training: $F_{1.16, 3.47} = 1.54$, $P = 0.30$; Neuron activity x Training: $F_{1.16, 3.48} = 1.42$, $P = 0.32$), log-transformed current interpress interval (l; 2-way ANOVA, Neuron activity: $F_{1, 3} = 11.71$, $P = 0.04$; Training: $F_{1.31, 3.94} = 0.38$, $P = 0.63$; Neuron activity x Training: $F_{1.31, 3.92} = 0.37$, $P = 0.63$), and the prior press interpress interval (s) (m; 2-way ANOVA, Neuron activity: $F_{1, 3} = 31.69$, $P = 0.01$; Training: $F_{1.07, 3.22} = 0.66$, $P = 0.48$; Neuron activity x Training: $F_{1.07, 3.21} = 0.62$, $P = 0.50$). R = correlation coefficient between actual and decoded press rate. **(n)** Visualization of representative single-subject, single-timepoint example population vectors with cosine distance between action-initiation and earned-reward neuronal population vectors. **(o-q)** Cosine angular distance between action-initiation and earned-reward D1⁺ neuronal population vectors across time and training (o), average (p; 1-way ANOVA, Training: $F_{1.17, 3.51} = 11.33$, $P = 0.03$, and maximum (q; peak; 1-way ANOVA, Training: $F_{1.06, 3.18} = 54.56$, $P = 0.004$) +/- 2.5 s peri-event. D1-cre: $N = 4$ (1 male). Data presented as mean ± s.e.m. Males = closed circles/solid lines, Females = open circles/dashed lines. *$P < 0.05$, **$P < 0.01$.

A subpopulation of DMS D1⁺ neurons stably encodes action initiation and information underlying behavioral strategy and develops action-outcome encoding.

DMS D1⁺ neurons are active around actions and earned rewards. Having found that the DMS D1⁺ neuronal population encodes the action-outcome association and other behavioral strategy information, we next asked how individual D1⁺ neurons encode task events during action-outcome learning and habit formation. A substantial proportion of D1⁺ neurons were significantly excited around, largely prior to, action initiation during each phase

of training (Figure 2a-b; see Extended Data Figure 2-1 for encoding of neurons from the whole population, see Extended Data Figure 2-2a-c for example classification). Another ensemble was inhibited by action initiation (Figure 2c; Extended Data Figure 2-3). Subpopulations of D1$^+$ neurons also encoded action termination (Extended Data Figure 2-4) and earned reward (Extended Data Figure 2-5). Thus, individual DMS D1$^+$ neurons encode actions and reward outcomes during learning and habit formation.

A subpopulation of DMS D1$^+$ neurons stably encodes action initiation and information underlying behavioral strategy. We next asked how the activity and encoding of individual DMS D1$^+$ neurons changes with action-outcome learning and habit formation. We classified neurons as significantly excited by action initiation during initial training. These *early action-initiation excited* D1$^+$ neurons were more active (Figure 2d-e) and modulated (Figure 2f; Extended Data Figure 2-2d) prior to the initiating lever press. They also encoded action initiation with high fidelity (Extended Data Figure 2-6). We looked prospectively to ask how this ensemble encoded action initiation as training progressed and found that these neurons continued to be active and modulated prior to action initiation throughout training and overtraining (Figure 2d-f; Extended Data Figure 2-2d). Further demonstrating ensemble stability, the activity around action initiation of both individual (Figure 2g) and the population (Figure 2h) of early action-initiation excited D1$^+$ neurons was significantly correlated across training phases. This stable D1$^+$ early action-initiation excited ensemble contained information both about ongoing actions and outcomes and variables related to goal-directed v. habitual behavioral control. Actions, checking for, and receiving reward, could all be decoded with a multi-class SVM trained on the activity of the D1$^+$ early action-initiation excited ensemble (Figure 2i). Decoding of action initiation improved with training, suggesting action-outcome learning improved the representation of action initiation in this ensemble. Behavioral engagement (press rate and switch rate; Figure 2j-k), current relative interpress interval (log interpress interval; Figure 2l), and past trial behavior (n-1 interpress interval; Figure 2m) could also be decoded with high accuracy across training using linear regression models trained on the activity of D1$^+$ early action-initiation excited neurons. To asses coding property stability, we used cross-session decoding. We trained a linear regression model to decode action rate from the activity of the early action-initiation excited D1$^+$ ensemble on the 1st training session. We could decode action rate not only for this early session, but also from the middle and overtraining training sessions (Figure 2n). Thus, an ensemble of DMS D1$^+$ neurons is excited prior to action initiation at the beginning of instrumental learning, continues to stably encode action initiation throughout training even as habits form, and stably conveys information relevant for both behavioral performance and strategy.

To provide converging evidence for DMS D1$^+$ neuron action-initiation encoding stability, we next took a retrospective approach. We classified D1$^+$ neurons significantly excited by action initiation during overtraining. These *'overtrain action-initiation excited'* neurons, too, were more active (Figure 2o-p) and modulated (Figure 2q; Extended Data Figure 2-2e) prior to action initiation. We looked retrospectively to ask whether this ensemble encoded action initiation at earlier training phases. It did. Overtrain action-initiation excited D1$^+$ neurons also showed elevated activity (Figure 2o-p) and modulation (Figure 2q) prior to action initiation during early and middle training. Action initiation encoding improved with training (Figure 2q; Extended Data Figure 2-2e). The activity around action initiation of individual (Figure 2r) and the population (Figure 2s) of overtrain action-initiation excited

D1$^+$ neurons was significantly correlated across training phases. Actions, checking for, and receiving reward, could all be decoded from D1$^+$ overtrain action-initiation excited ensemble activity (Figure 2t), as could behavioral engagement (Figure 2u-v), log interpress interval (Figure 2w), and information about past trial action performance (Figure 2x). Encoding properties were also stable. When we trained a linear regression model to decode action rate from the activity of overtrain action-initiation excited D1$^+$ neurons on the last training session, we could reliably decode action rate not only from this session, but also the preceding training sessions (Figure 2y). Thus, regardless of whether we identified neurons early in training and looked prospectively or during overtraining and looked retrospectively, we found that about half (Figure 2z-aa) of the DMS D1$^+$ action-initiation excited neurons stably encoded action initiation and conveyed information relevant for both behavioral performance and strategy throughout action-outcome learning and habit formation. The D1$^+$ neuron ensemble that was inhibited around action initiation was similarly stable (Extended Data Figure 2-3).

Stable DMS D1$^+$ action-initiation neurons develop encoding of the action-outcome relationship with learning. If activity of the stable DMS D1$^+$ action-initiation ensemble relates to action-outcome learning, we reasoned that it might not solely encode action initiation, but, with learning, might come to encode the action-outcome relationship. Indeed, we found that stable action-initiation excited neurons also came to encode action termination and earned reward outcomes with learning (Figure 2ab-ac; Extended Data Figure 2-7). The stable action-initiation excited D1$^+$ neurons showed action bracketing[66-68], being excited before the start and end of an action bout. The percentage of stable action-initiation excited neurons encoding both action initiation and termination grew from 41.3% to 87.03% with training. Supporting representation of the action-outcome association, the proportion of the stable action-initiation neurons that also encoded earned reward grew from 18.7% to 63.9%. Thus, a subpopulation of DMS D1$^+$ neurons stably encoded actions and, with learning, encoded the action-outcome relationship. Interestingly, the D1$^+$ subpopulation encoding earned reward was not stable across training (Extended Data Figure 2-5). Most of the D1$^+$ neurons that did stably encode the earned reward and all of those that gained reward encoding with training also encoded action initiation (Figure 2-5m-p). Those neurons that stopped encoding the reward with training were less likely to encode action initiation. This suggests the D1$^+$ neuronal representation of earned reward outcomes merged into that of the action. Thus, a subpopulation of DMS D1$^+$ neurons stably encodes actions and, with learning, comes to represent the action-outcome relationship that supports agency.

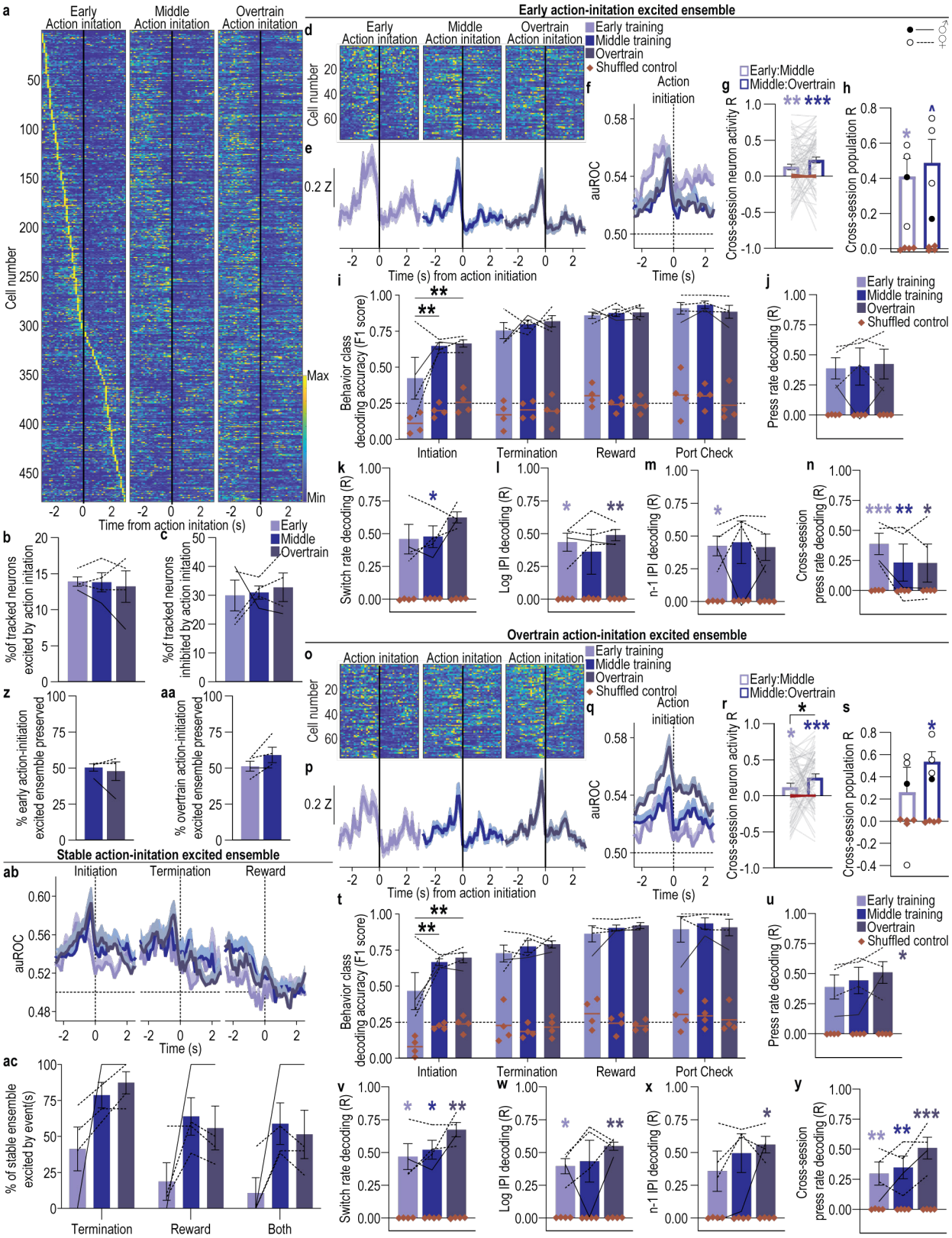


Figure 2: A subpopulation of DMS D1⁺ neurons stably encodes action initiation and information underlying behavioral strategy and develops action-outcome encoding. (a) Heat map of minimum to maximum deconvolved activity (sorted by total activity) of each coregistered DMS D1⁺ neuron around lever-press action initiation. **(b-c)** Percent of coregistered neurons (Average 137 coregistered neurons/mouse, s.e.m. 37.73) significantly excited (b; 1-way ANOVA, $F_{1.13, 3.63} = 0.14$, $P = 0.76$), or inhibited (c; 1-way ANOVA, $F_{1.29, 3.86} = 0.14$, $P = 0.79$) around action initiation. **(d-f)** Activity and modulation across training of D1⁺ early action-initiation excited neurons (N = 77 neurons/4 mice; average 19.25 neurons/mouse, s.e.m. = 5.36). Heat map (d), Z-scored activity (e), and area under the receiver operating characteristic curve (auROC) modulation index (f) of early action-initiation excited neurons around action initiation across training. **(g-h)** Cross-session correlation of the activity around action initiation of each early action-initiation excited neuron (g; 2-way ANCOVA, Neuron activity: $F_{1, 75} = 22.92$, $P < 0.001$; Training: $F_{1, 75} = 0.40$, $P = 0.53$; Activity x Training: $F_{1, 75} = 0.24$, $P = 0.63$) or the

population activity of these neurons (h; 2-way ANOVA, Neuron activity: $F_{1, 3} = 39.10$, $P = 0.008$; Training: $F_{1, 3} = 0.70$, $P = 0.18$; Activity x Training: $F_{1, 3} = 0.16$, $P = 0.72$). **(i)** Behavior class decoding accuracy from D1⁺ early action-initiation excited neuron activity in 0.5-s period prior to event compared to shuffled control. Line at 0.25 = chance. 3-way ANOVA, Neuron activity (v. shuffled): $F_{1, 3} = 1118.27$, $P < 0.001$; Training: $F_{2, 6} = 2.79$, $P = 0.14$; Behavior class: $F_{3, 9} = 16.94$, $P < 0.001$; Neuron activity x Training: $F_{2, 6} = 2.10$, $P = 0.20$; Neuron activity x Behavior: $F_{3, 9} = 3.53$, $P = 0.06$; Training x Behavior: $F_{6, 18} = 3.29$, $P = 0.23$; Neuron activity x Training x Behavior: $F_{6, 18} = 0.30$, $P = 0.93$. **(j-m)** Linear decoding accuracy from D1⁺ early action-initiation excited neuron activity of press rate (j; 2-way ANOVA, Neuron activity: $F_{1, 3} = 12.70$, $P = 0.04$; Training: $F_{1.16, 3.49} = 0.08$, $P = 0.83$; Neuron activity x Training: $F_{1.20, 3.60} = 0.09$, $P = 0.82$), switch rate (k; 2-way ANOVA, Neuron activity: $F_{1, 3} = 68.28$, $P = 0.004$; Training: $F_{1.40, 4.20} = 1.79$, $P = 0.26$; Neuron activity x Training: $F_{1.38, 4.13} = 1.67$, $P = 0.28$), log-transformed current interpress interval; 2-way ANOVA, Neuron activity: $F_{1, 3} = 31.18$, $P = 0.01$; Training: $F_{1.62, 4.85} = 3.10$, $P = 0.14$; Neuron activity x Training: $F_{1.63, 4.84} = 2.99$, $P = 0.14$), and the prior press interpress interval (s) (m; 2-way ANOVA, Neuron activity: $F_{1, 3} = 21.31$, $P = 0.02$; Training: $F_{1.06, 3.18} = 0.06$, $P = 0.83$; Neuron activity x Training: $F_{1.06, 3.71} = 0.03$, $P = 0.88$). R = correlation coefficient between actual and decoded press rate. **(n)** Cross-session decoding accuracy of press rate from the activity of D1⁺ early action-initiation-excited neuron population activity on the 1st training session. Planned, Bonferroni corrected, 2-tailed t-tests, Early: $t_6 = 7.94$, $P = 0.0006$, 95% CI 0.23 - 0.55; Middle: $t_6 = 4.83$, $P = 0.009$, 95% CI 0.08 - 0.40; Overtrain: $t_6 = 4.68$, $P = 0.01$, 95% CI 0.07 - 0.39. **(o-q)** Activity and modulation across training of D1⁺ overtrain action-initiation excited neurons ($N = 76$ neurons/4 mice; average 19 neurons/mouse, s.e.m. 5.49). Heat map (o), Z-scored activity (p), and auROC modulation index (q) of overtrain action-initiation excited neurons around action initiation across training. **(r-s)** Cross-session correlation of the activity around action initiation of each overtrain action-initiation excited neuron (r; 2-way ANCOVA, Neuron activity: $F_{1, 74} = 21.06$, $P < 0.001$; Training: $F_{1, 74} = 0.23$, $P = 0.64$; Activity x Time: $F_{1, 74} = 0.26$, $P = 0.61$) or the population activity of these neurons (s; 2-way ANOVA, Neuron activity: $F_{1, 3} = 8.25$, $P = 0.06$; Training: $F_{1, 3} = 1.76$, $P = 0.28$; Activity x Time: $F_{1, 3} = 2.29$, $P = 0.23$). **(t)** Behavior class decoding accuracy from D1⁺ overtrain action-initiation excited neuron activity. 3-way ANOVA, Neuron activity: $F_{1, 3} = 262.75$, $P < 0.001$; Training: $F_{2, 6} = 2.78$, $P = 0.14$; Behavior $F_{3, 9} = 11.31$, $P = 0.002$; Neuron activity x Training: $F_{2, 6} = 2.16$, $P = 0.20$; Neuron activity x Behavior: $F_{3, 9} = 3.86$, $P = 0.05$; Training x Behavior: $F_{6, 18} = 3.27$, $P = 0.025$; Neuron activity x Training x Behavior: $F_{6, 18} = 0.52$, $P = 0.79$. **(u-x)** Linear decoding accuracy from D1⁺ overtrain action-initiation excited neuron activity of press rate (u; 2-way ANOVA, Neuron activity: $F_{1, 3} = 24.76$, $P = 0.02$; Training: $F_{1.02, 3.05} = 1.25$, $P = 0.35$; Population activity x Training: $F_{1.00, 3.04} = 1.24$, $P = 0.35$), switch (v; 2-way ANOVA, Neuron activity: $F_{1, 3} = 86.87$, $P = 0.003$; Training: $F_{1.62, 4.85} = 3.10$, $P = 0.14$; Neuron activity x Training: $F_{1.61, 4.84} = 2.99$, $P = 0.14$), log-transformed current interpress interval (w; 2-way ANOVA, Neuron activity: $F_{1, 3} = 44.96$, $P = 0.007$; Training: $F_{1.41, 4.23} = 0.80$, $P = 0.46$; Neuron activity x Training: $F_{1.41, 4.23} = 0.81$, $P = 0.46$), and the prior press interpress interval (s) (x; 2-way ANOVA, Neuron activity: $F_{1, 3} = 20.22$, $P = 0.02$; Training: $F_{1.39, 4.17} = 1.25$, $P = 0.35$; Neuron activity x Training: $F_{1.39, 4.17} = 1.27$, $P = 0.35$). **(y)** Cross-session decoding accuracy of lever-press rate from the activity of D1⁺ overtrain action-initiation-excited neuron population activity on the 8th training session. Planned, Bonferroni corrected, 2-tailed t-tests, Early: $t_6 = 3.64$, $P = 0.005$, 95% CI 0.12 to 0.47; Middle: $t_6 = 6.48$, $P = 0.002$, 95% CI 0.17 - 0.53; Overtrain: $t_6 = 9.49$, $P = 0.002$, 95% CI 0.33 - 0.69. **(z)** Percent of D1⁺ early action-initiation excited neurons that continued to be significantly excited by action initiation on the 4th and 8th training sessions. 2-tailed Wilcoxon signed rank test, W = -1.00, P > 0.99. **(aa)** Percent of D1⁺ overtrain action-initiation excited neurons that were also significantly excited by action initiation on 1st and 4th training sessions. 2-tailed t-test, $t_3 = 2.02$, $P = 0.14$, 95% CI -4.52 - 20.15. **(ab)** Modulation around action initiation, termination, and earned reward across training of D1⁺ stable action-initiation excited neurons ($N = 31$ neurons/4 mice; average 7.75 neurons/mouse, s.e.m. = 2.56). **(ac)** Percent of coregistered D1⁺ neurons stably excited by action initiation across training that were also significantly modulated by action termination and reward collection. 2-way ANOVA, Event: $F_{1, 11} = 7.81$, $P = 0.06$; Training: $F_{1.02, 3.05} = 4.37$, $P = 0.13$; Event x Training: $F_{1.86, 5.59} = 1.76$, $P = 0.25$. D1-cre: $N = 4$ (1 male). Data presented as mean ± s.e.m. Males = closed circles/solid lines, Females = open circles/dashed lines. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

DMS D1⁺ neurons support action-outcome learning and agency.

DMS D1⁺ neuronal activity is necessary for the action-outcome learning underlying agency. Given that DMS D1⁺ neurons encode the action-outcome relationship and convey information related to the extent of goal-directed v. habitual behavioral control, we next reasoned they might mediate action-outcome learning and, therefore, agency. To test this, we chemogenetically inhibited DMS D1⁺ neurons during learning. We chose chemogenetic manipulation because DMS D1⁺ neurons were endogenously active at multiple time points (action initiation, termination, and reward outcome). Such manipulations also avoid artificially synchronizing subpopulations of neurons with different activity profiles and avoid the reinforcing/punishing[69,70] and locomotor[71-73] effects that can occur with optogenetic manipulations of these cells. We expressed the inhibitory designer receptor human M4 muscarinic receptor (hM4Di) selectively in DMS D1⁺ neurons of DRD1a-cre or wildtype control mice (Figure 3a-b). Mice were trained to lever press to earn food-pellet rewards on a random-interval reinforcement schedule (Figure 3c). We gave mice limited (4 sessions) training to preserve dominance of goal-directed behavioral control in wildtype controls (Extended Data Figure 1-1). Prior to each training session, mice received the hM4Di ligand clozapine-N-oxide (CNO; 2.0 mg/kg[74-76] i.p.) to, in hM4Di-expressing DRD1a-cre subjects, inactivate DMS D1⁺ neurons (see Extended Data Figure 3-1 for validation). Chemogenetic inactivation of DMS D1⁺ neurons did not

significantly alter acquisition of the instrumental behavior (Figure 3d; see Extended Data Figure 3-2 for food-port entry and presses/reward). Thus, DMS D1$^+$ neuronal activity is not necessary for general acquisition or execution of instrumental behavior. Inhibition of D1$^+$ neurons did slightly decrease behavioral switch rate and significantly decreased interpress interval consistency (Extended Data Figure 3-3), suggestive of reduced goal-directed behavioral control. To directly evaluate behavioral strategy, training was followed by the outcome-specific devaluation test. No CNO was given on test. If subjects have learned the action-outcome relationship and are using this to support prospective consideration of action consequences for flexible, goal-directed decision making, they will reduce lever pressing when the outcome is devalued. We saw such agency in controls (Figure 3e-f). D1$^+$ neuron inhibition disrupted the action-outcome learning underlying goal-directed decision making as evidenced by subsequent insensitivity to outcome devaluation (Figure 3e-f). Inhibition of D1$^+$ neurons during learning prevented the action-outcome learning needed for flexible, goal-directed decision making, forcing instrumental behavior to rely on an inflexible habitual strategy. Thus, DMS D1$^+$ neurons are active during instrumental learning and this activity mediates the action-outcome learning that supports flexible, goal-directed decision making.

DMS D1$^+$ neuronal activity is sufficient to promote goal-directed behavioral control. Because DMS D1$^+$ neurons encode the action-outcome association and other information underlying behavioral strategy, even after habits have formed, we reasoned that these neurons might drive the agency that relies on this knowledge. If this is true, then augmenting D1$^+$ neuron activity should promote agency, leading to continued flexible goal-directed behavioral control even after overtraining. To test this, we chemogenetically activated D1$^+$ neurons during learning and overtraining and then probed behavioral strategy using the devaluation test. We expressed the excitatory designer receptor human M3 muscarinic receptor (hM3Dq) or fluorophore control selectively in DMS D1$^+$ neurons of DRD1a-cre or wildtype control mice (Figure 3g-h). Mice were trained to lever press to earn food-pellet rewards (Figure 3i) and were overtrained (8 sessions) to promote habit formation in controls (Extended Data Figure 1-1). Prior to each training session, mice received the hM3Dq ligand CNO (0.2 mg/kg[77-80] i.p.) to, in hM3Dq-expressing DRD1a-cre subjects, activate D1$^+$ neurons (Extended Data Figure 3-1). Chemogenetic activation of DMS D1$^+$ neurons did not alter acquisition of the instrumental behavior (Figure 3j; see also Extended Data Figure 3-3d-f), but did promote goal-directed behavioral control. Whereas controls showed evidence of habit formation, insensitivity to devaluation, subjects for which we activated D1$^+$ neurons during learning were sensitive to devaluation, indicating the dominance of goal-directed decision making (Figure 3k-l). Thus, enhanced DMS D1$^+$ neuronal activity promotes dominance of flexible, goal-directed decision making over inflexible habitual behavioral control.

DMS D1$^+$ neuronal activity mediates the expression of goal-directed decision making. Given evidence that DMS D1$^+$ neurons encode the action-outcome association and information related to the extent of goal-directed v. habitual control, we next reasoned that these neurons may also support the online use of agency for flexible goal-directed decision making. If this is true, then D1$^+$ neuronal activity should be required to express goal-directed decision making by adapting when the outcome value changes. To test this, we again chemogenetically inhibited D1$^+$ neurons, this time at test rather than during learning. We expressed hM4Di or fluorophore control selectively

in DMS D1⁺ neurons of DRD1a-cre or wildtype control mice (Figure 3m-n). Mice received limited training to lever press to earn food-pellet rewards (Figure 3o). All mice acquired the instrumental behavior (Figure 3p). Mice then received the outcome-specific devaluation test. After sensory-specific satiety, prior to the lever-pressing probe test mice received CNO (2.0 mg/kg i.p.) to inactivate D1⁺ neurons. Controls showed flexible, goal-directed decision making, reducing action performance when the outcome was devalued (Figure 3q-r). Chemogenetic D1⁺ neuron inhibition disrupted the expression of such agency, as evidenced by insensitivity to devaluation (Figure 3q-r). Inhibition of D1⁺ neurons prevented subjects from acting with agency and instead forced instrumental behavior to rely on the inflexible habit system. Thus, DMS D1⁺ neurons mediate both action-outcome learning and the application of this learned agency for flexible, goal-directed decision making.

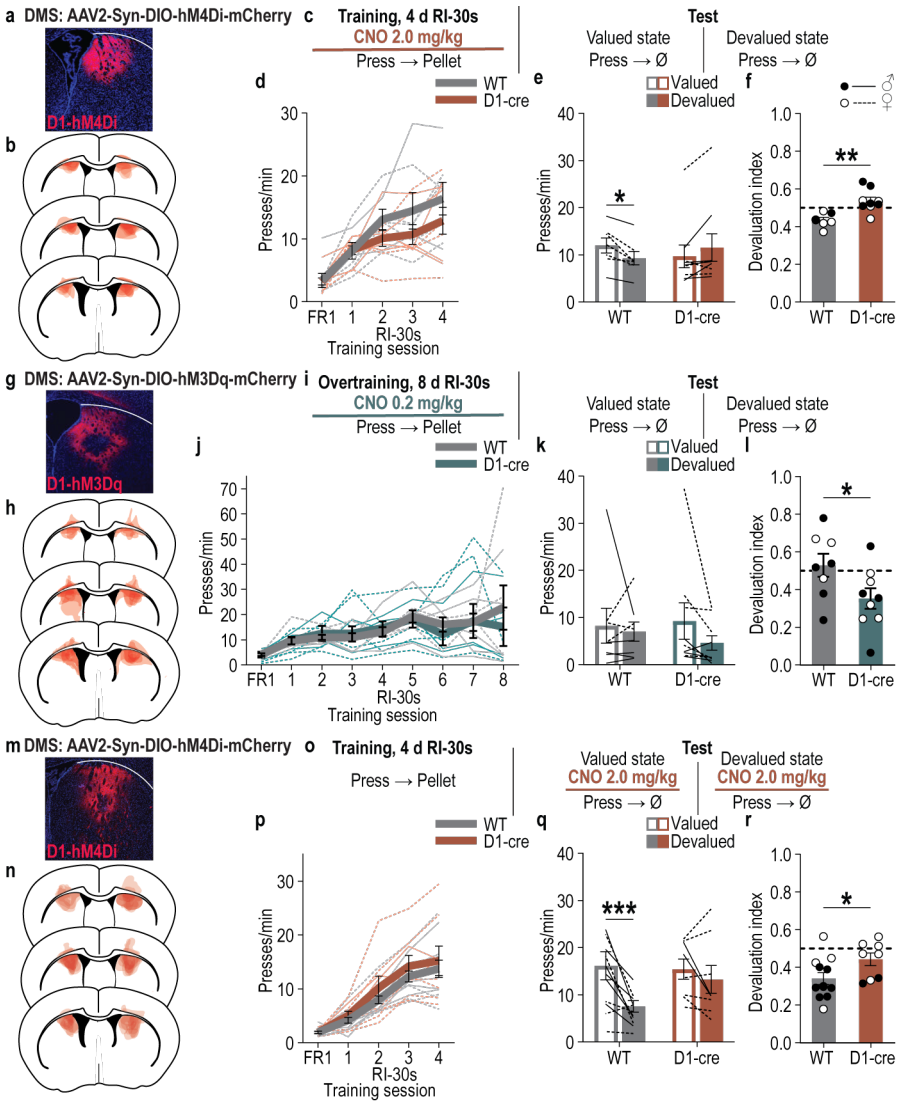


Figure 3: DMS D1⁺ neurons support action-outcome learning and goal-directed decision making. (a-f) Chemogenetic inactivation of DMS D1⁺ neurons during learning. WT: $N = 7$ (2 males); D1-cre: $N = 9$ (5 males). **(a)** Representative immunofluorescent image of cre-dependent hM4Di expression in DMS. **(b)** Map of DMS cre-dependent hM4Di expression for all subjects. **(c)** Procedure. RI, random-interval reinforcement schedule; CNO, clozapine-N-oxide. **(d)** Training press rate. 2-way ANOVA, Training: $F_{2.22, 31.01} = 27.65$, $P < 0.0001$; Genotype: $F_{1, 14} = 1.27$, $P = 0.28$; Training x Genotype: $F_{4, 56} = 1.24$, $P = 0.30$. **(e)** Test press rate. 2-way ANOVA, Value x Genotype: $F_{1, 14} = 14.30$, $P = 0.002$; Value: $F_{1, 14} = 0.49$, $P = 0.49$; Genotype: $F_{1, 14} = 0.00005$, $P = 0.99$. **(f)** Devaluation index [(Devalued condition presses)/(Valued condition presses + Devalued presses)]. 2-tailed t-test, $t_{14} = 4.01$; $P = 0.001$, 95% CI -0.16 - -0.05. **(g-l)** Chemogenetic activation of D1⁺ neurons during overtraining. WT: $N = 6$ (4 males); D1-cre: $N = 6$ (3 males). **(g)** Representative immunofluorescent image of cre-dependent hM3Dq expression in DMS. **(h)** Map of DMS cre-dependent hM3Dq expression. **(i)** Procedure. **(j)** Training press rate. 2-way ANOVA, Training: $F_{2.44, 24.43} = 3.23$, $P = 0.048$; Genotype: $F_{1, 10} = 0.31$, $P = 0.59$; Training x Genotype: $F_{8, 80} = 0.68$, $P = 0.71$. **(k)** Test press rate. 2-way ANOVA, Value x Genotype: $F_{1, 10} = 3.37$, $P = 0.10$; Value: $F_{1, 10} = 0.56$, $P = 0.47$; Genotype: $F_{1, 10} = 0.30$, $P = 0.59$.

(l) Devaluation index. 2-tailed t-test, t_{10} = 3.07; P = 0.01, 95% CI -0.45 - -0.07. **(m-r)** Chemogenetic inactivation of D1$^+$ neurons during test of behavioral strategy after learning. WT: N = 12 (7 males); D1-cre: N = 12 (5 males). **(m)** Representative immunofluorescent image of cre-dependent hM4Di expression in DMS. **(n)** Map of DMS cre-dependent hM4Di expression. **(o)** Procedure. **(p)** Training press rate. 2-way ANOVA, Training: $F_{1.59, 28.63}$ = 68.95, P < 0.0001; Genotype: $F_{1, 18}$ = 0.74, P = 0.40; Training x Genotype: $F_{4, 72}$ = 0.44, P = 0.78. **(q)** Test press rate. 2-way ANOVA, Value x Genotype: $F_{1, 18}$ = 4.23, P = 0.05; Value: $F_{1, 18}$ = 11.93, P = 0.003; Genotype: $F_{1, 18}$ = 0.64, P = 0.44. **(r)** Devaluation index. 2-tailed t-test, t_{18} = 2.15; P = 0.045, 95% CI 0.003 - 0.20. Data presented as mean ± s.e.m. Males = closed circles/solid lines, Females = open circles/dashed lines. *P < 0.05, **P < 0.01, ***P < 0.001.

The DMS A2A$^+$ neuronal population encodes behavior and information underlying behavioral strategy.

We next asked how DMS A2A$^+$ neuronal activity relates to action-outcome learning and habit formation using cellular-resolution, microendoscopic imaging of jGCaMP7s selectively expressed in A2A$^+$ neurons of A2A-cre mice (Figure 4a-d). As with D1$^+$ neurons, we recorded calcium activity in DMS A2A$^+$ neurons as mice learned to lever press to earn food-pellet rewards on a random-interval reinforcement schedule and then probed behavioral strategy with an outcome-specific devaluation test (Figure 4e). All mice acquired the instrumental behavior (Figure 4f; see Extended Data Figure 4-1 for food-port entry data). Following overtraining, half the mice were insensitive to devaluation (Figure 4g-h), indicating routine habit formation. Unexpectedly, half the A2A-cre mice did not form habits with overtraining (Extended Data Figure 1-4). To capture activity related to action-outcome learning and habit formation, we focused on the subjects that formed habits and analyzed and coregistered (Figure 4d; Extended Data Figure 4-2) calcium activity across the beginning, middle, and end of training (Supplemental Table 1).

Instrumental actions, reward outcomes, and information underlying behavioral strategy can be decoded from DMS A2A$^+$ neuronal population activity. We again used decoding to probe the information content available in the activity of the tracked A2A$^+$ neuronal population. Using a multi-class SVM algorithm, we first found that behavioral events (action initiation, action termination, non-reinforced food-port checks, and earned reward consumption) could be reliably decoded from A2A$^+$ neuronal activity (Figure 4i). Decoding of action initiation initially improved with learning, but decreased with overtraining suggesting fluctuation in action coding with action-outcome learning and habit formation. Using linear regression models, we next found that both behavioral engagement and organizational variables related to goal-directed v. habitual behavioral strategy could be decoded from the A2A$^+$ neuronal population. Press rate, switch rate, log interpress interval, and past trial interpress interval could all be decoded from the A2A$^+$ neuronal population activity (Figure 4j-m; see Extended Data Figure 4-3 for example of decoded behavior). Thus, the DMS A2A$^+$ neuronal population can convey information about actions, outcomes, and behavioral strategy.

The DMS A2A$^+$ neuronal population does not represent the action-outcome association. We next used a population vector analysis to ask whether, with learning, the DMS A2A$^+$ neuronal population comes to encode the action-outcome association underlying agency. We found it does not. The cosine distance between the population vectors for action initiation and reward did not significantly change with training (Figure 4n-q). A representational similarly analysis also found no evidence the A2A$^+$ neuronal population action and outcome representation becomes more similar with learning (Extended Data Figure 4-4). Thus, the DMS A2A$^+$ neuronal population does not represent the action-outcome association.

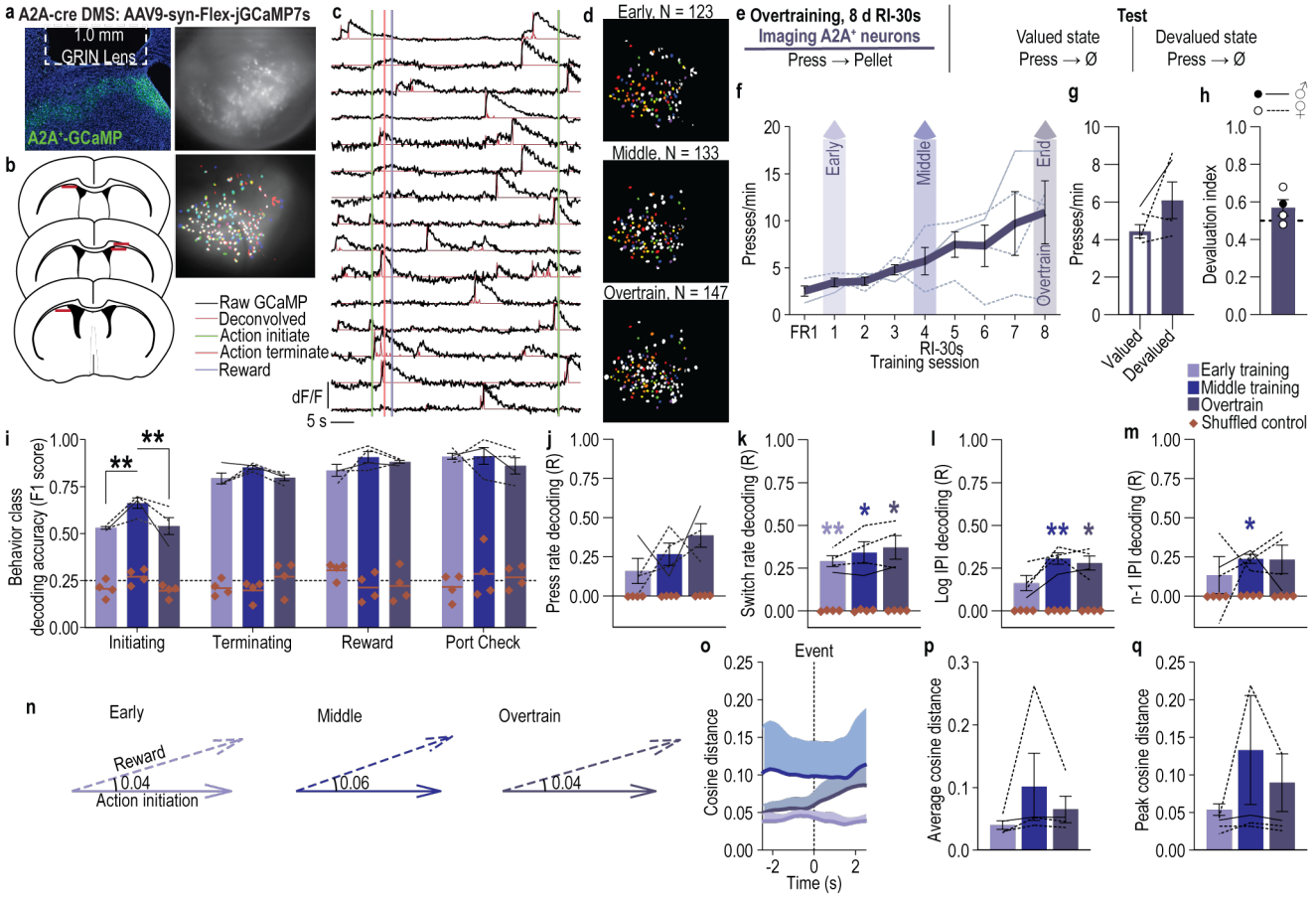


Figure 4: The DMS A2A⁺ neuronal population encodes behavior and information underlying behavioral strategy. (a) Representative image of cre-dependent jGCaMP7s expression in DMS A2A⁺ neurons (top left) and maximum intensity projection of the field-of-view with (top right) and without (bottom right) demarcation of cell ROIs for one session. **(b)** Map of DMS GRIN lens placements. **(c)** Representative dF/F and deconvolved calcium signal. **(d)** Representative recorded D1⁺ neuron spatial footprints during the first (early), 4th (middle), and 8th (overtrain) random-interval training sessions. Colored, co-registered neurons; white, non-co-registered neurons. **(e)** Procedure. RI, random-interval reinforcement schedule. **(f)** Training press rate. 1-way ANOVA, Training: $F_{1.39, 4.18} = 3.56$, $P = 0.12$. **(g)** Test press rate. 2-tailed t-test, $t_3 = 1.50$, $P = 0.23$, 95% CI -1.86 - 5.16. **(h)** Devaluation index [(Devalued condition presses)/(Valued condition presses + Devalued presses)]. One-tailed Bayes factor, $BF_{10} = 0.19$. **(i)** Behavior class decoding accuracy from A2A⁺ coregistered neuron activity in 0.5-s period prior to event compared to shuffled control. Line at 0.25 = chance. 3-way ANOVA, Neuron activity (v. shuffled): $F_{1, 3} = 1570.57$, $P < 0.001$; Training session: $F_{2, 6} = 5.14$, $P = 0.05$; Behavior class: $F_{3, 9} = 23.33$, $P < 0.001$; Neuron activity x Training: $F_{2, 6} = 3.72$, $P = 0.09$; Neuron activity x Behavior: $F_{3, 9} = 17.65$, $P < 0.001$; Training x Behavior: $F_{6, 18} = 2.00$, $P = 0.12$; Neuron activity x Training x Behavior: $F_{6, 18} = 1.66$, $P = 0.19$. **(j-m)** Linear decoding accuracy from A2A⁺ coregistered neuron activity of press rate (j; 2-way ANOVA, Neuron activity: $F_{1, 3} = 72.91$, $P = 0.003$; Training: $F_{1.33, 4.00} = 1.91$, $P = 0.25$; Neuron activity x Training: $F_{1.35, 4.04} = 1.74$, $P = 0.27$), switch rate (k; 2-way ANOVA, Neuron activity: $F_{1, 3} = 37.30$, $P = 0.009$; Training: $F_{1.36, 4.08} = 2.96$, $P = 0.16$; Neuron activity x Training: $F_{1.38, 4.12} = 2.85$, $P = 0.17$), log-transformed current interpress interval (l; 2-way ANOVA, Neuron activity: $F_{1, 3} = 59.49$, $P = 0.005$; Training: $F_{1.99, 5.96} = 6.84$, $P = 0.03$; Neuron activity x Training: $F_{1.99, 5.96} = 6.83$, $P = 0.03$), and the prior press interpress interval (s) (m; 2-way ANOVA, Neuron activity: $F_{1, 3} = 106.40$, $P = 0.02$; Training: $F_{1.26, 3.78} = 0.64$, $P = 0.83$; Neuron activity x Training: $F_{1.27, 3.81} = 0.30$, $P = 0.67$). R = correlation coefficient between actual and decoded press rate. **(n)** Visualization of representative single-subject, single-timepoint example population vectors with cosine distance between action-initiation ("action") and earned-reward ("reward") neuronal population vectors. **(o-q)** Cosine angular distance between action-initiation and earned-reward A2A⁺ neuronal population vectors across time and training centered around each event (o), average (p; 1-way ANOVA, Training: $F_{1.00, 3.00} = 1.57$, $P = 0.30$), and maximum (q; peak; 1-way ANOVA, Training: $F_{1.00, 3.02} = 1.37$, $P = 0.33$) +/- 2.5 s around each event. Data presented as mean ± s.e.m. A2A-cre: $N = 4$ (1 male). Males = closed circles/solid lines, Females = open circles/dashed lines. *$P < 0.05$, **$P < 0.01$.

The subpopulation of DMS A2A⁺ neurons that encodes action initiation reorganizes with habit formation.

DMS A2A⁺ neurons are active around actions. We next asked how individual A2A⁺ neurons encode task events during action-outcome learning and the transition to habit. Whether looking at the entire population or the neurons we could track across training, a substantial proportion of A2A⁺ neurons were significantly excited around, largely prior to, action initiation during each phase of training (Figure 5a-b; see Extended Data Figure 5-1 for encoding in whole population; see Extended Data Figure 5-2a-c for an example classification). Another ensemble of A2A⁺

neurons was inhibited by action initiation (Figure 5c; Extended Data Figure 5-3). Only a few tracked A2A$^+$ neurons were excited by action termination (Extended Data Figure 5-4) or reward (Extended Data Figure 5-5), though substantial proportions were inhibited by these events (Extended Data Figures 5-1). Indeed, A2A$^+$ neurons appeared to quiet their activity during consumption of the earned reward (Extended Data Figure 5-1). Thus, DMS A2A$^+$ neurons encode actions during learning and as habits form with overtraining.

We next asked how the activity and encoding of individual DMS A2A$^+$ neurons changes with action-outcome learning and habit formation. We classified neurons that were significantly excited by action initiation early in training. These early action-initiation excited A2A$^+$ neurons were more active (Figure 5d-e) and modulated (Figure 5f; Extended Data Figure 5-2d) prior to action initiation than after and encoded action initiation with moderate fidelity (Extended Data Figure 5-6). The early A2A$^+$ action-initiation excited neurons lost their encoding of action initiation with training (Figure 5f; Extended Data Figure 5-2d). Neither the activity around action initiation of individual (Figure 5g), nor the population (Figure 5h) of early action-initiation excited A2A$^+$ neurons was significantly correlated across training. Thus, early action-initiation excited A2A$^+$ neurons only transiently encoded actions. Behavioral events (actions, checking for, and receiving reward) could all be decoded with a multi-class SVM trained on the A2A$^+$ early action-initiation excited neuron activity (Figure 5i). Neither press rate (Figure 5j), nor information about past action performance (past trial interpress interval; Figure 5m) could be decoded with linear regression models trained on the activity of the A2A$^+$ early action-initiation excited neurons, but behavioral switch rate (Figure 2k) and log interpress interval (Figure 5l) could be decoded. Thus, A2A$^+$ early action-initiation excited neurons convey information about behavioral transitions and relative performance, but not engagement or recent past behavior. Demonstrating coding-property instability, we were unable to decode action rate across sessions (Figure 5n). Therefore, early action-initiation excited DMS A2A$^+$ neurons convey behavioral performance and some information relevant for strategy, especially behavioral transitions, but only transiently encode actions, losing their encoding as habits form with overtraining. These neurons did not retune to become excited by other task events, though some became inhibited by actions or rewards (Extended Data Figure 5-7).

Although the early action-initiation excited A2A$^+$ neurons stopped encoding actions, at overtraining there were neurons excited by action initiation. We classified these overtrain action-initiation excited A2A$^+$ neurons and found that they, too, were more active (Figure 5o-p) and modulated (Figure 5q, Extended Data Figure 5-2e) prior to action initiation. We looked retrospectively and found that activity around and modulation by action initiation of this A2A$^+$ ensemble grew with training and overtraining (Figure 5o-q; Extended Data Figure 5-2e). We did not find significant cross-session correlations between the activity of individual or the population of overtrain action-initiation excited A2A$^+$ neurons (Figure 5r-s). Thus, as habits form with overtraining a new subpopulation of DMS A2A$^+$ neurons came to encode action initiation. This did not result from retuning. Overtrain action-initiation A2A$^+$ neurons were largely not excited by other events earlier in training, though, a small proportion were inhibited by actions at earlier task phases (Extended Data Figure 5-7). Behavioral events could be decoded from the activity of the A2A$^+$ overtrain action-initiation excited ensemble (Figure 5t), but none of the linear variables related to behavioral strategy could be decoded (Figure 5u-x), and we were unable to decode

action rate across sessions from these neurons (Figure 5y). Thus, the neurons that pickup action-initiation encoding with overtraining convey information about behavioral performance, but not strategy.

Together these data indicate that one subpopulation of A2A$^+$ neurons transiently encodes actions during early action-outcome learning and another slowly starts to encode actions as routine habits form. Indeed, after overtraining only about 15% of the early action-initiation excited A2A$^+$ neuron ensemble continued to significantly encode action initiation (Figure 5z) and only about 19% of the overtrain action-initiation A2A$^+$ ensemble significantly encoded action initiation at earlier training phases (Figure 5aa). This reorganization contrasts with the D1$^+$ neurons (Figure 5ab).

The DMS A2A$^+$ ensemble that was inhibited around action initiation was similarly unstable (Extended Data Figure 5-3), but did contain information related to behavioral strategy. Action rate (Figure 5ad-ae) and behavioral strategy variables (Extended Data Figure 5-3) could be decoded from the activity of the overtrain action-initiation inhibited neurons. Thus, although not stable in their encoding, DMS A2A$^+$ neurons that are inhibited by action convey goal-directed v. habitual behavioral control information.

DMS A2A$^+$ action encoding is more stable if habits do not form. We next asked whether habit formation contributes to the reorganization of DMS A2A$^+$ neuronal action-initiation encoding by exploiting the serendipity that half the A2A-cre subjects did not form habits with overtraining. Like subjects that formed habits, a substantial proportion of A2A$^+$ neurons were activated around, largely prior to, action initiation in subjects that did not form habits (Extended Data Figure 5-8). In mice that did not form habits, the A2A$^+$ neurons maintained elevated responses to action initiation across training (Extended Data Figure 5-8). The stability of the early action-initiation excited A2A$^+$ neurons, assessed by the extent to which it continued to be excited prior to action initiation after overtraining, negatively correlated with devaluation index (Figure 5ac). More stability was associated with greater sensitivity to devaluation and, thus, stronger goal-directed behavioral control. Thus, habit formation contributes to DMS A2A$^+$ neuronal reorganization of action encoding.

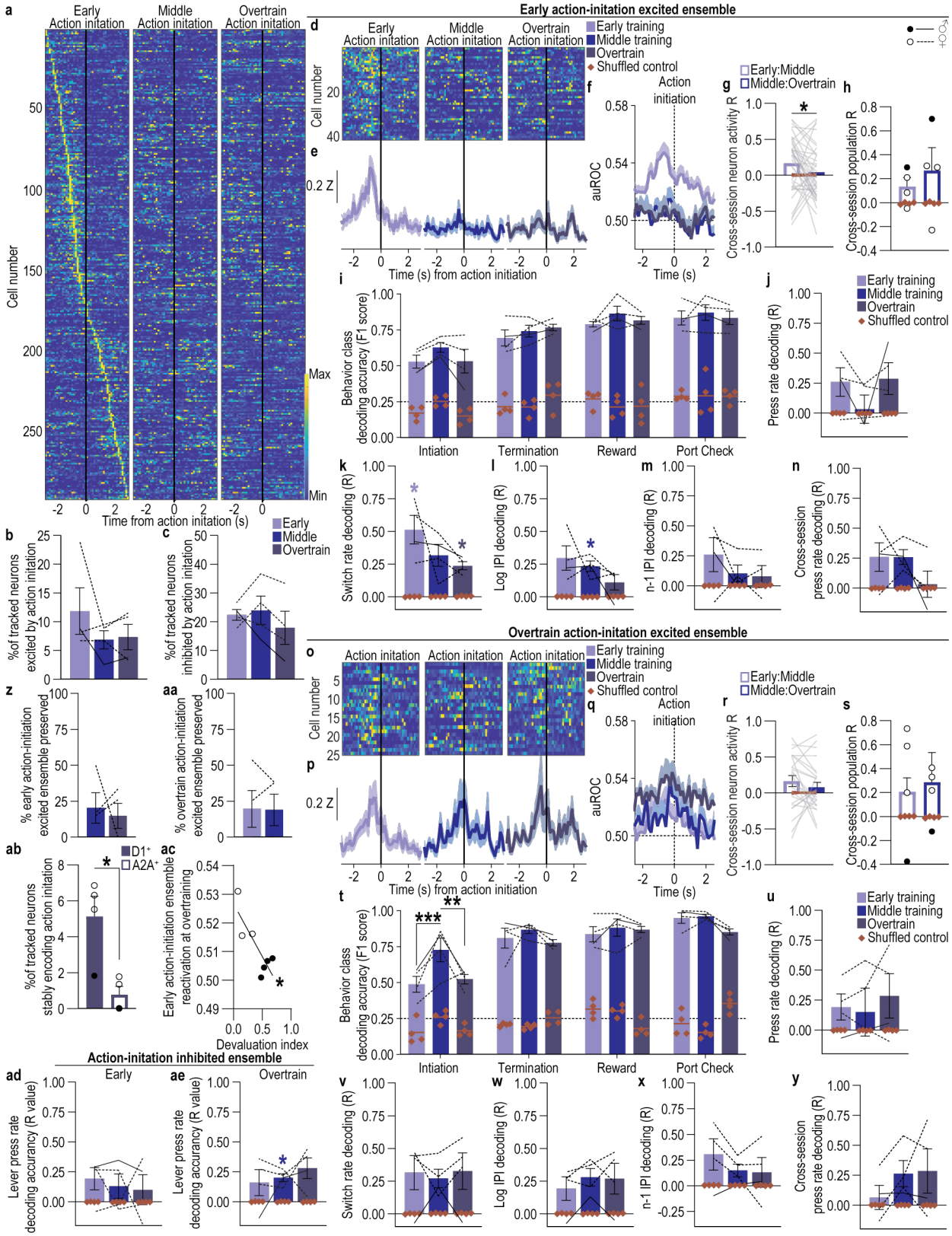


Figure 5: The subpopulation of DMS A2A⁺ neurons that encodes action initiation reorganizes with habit formation. (a) Heat map of minimum to maximum deconvolved activity (sorted by total activity) of each coregistered DMS A2A⁺ neuron around lever-press action initiation. **(b-c)** Percent of coregistered neurons (Average 73.25 coregistered neurons/mouse, s.e.m. 17.09) significantly excited (b; 1-way ANOVA, $F_{1.16, 3.48} = 1.79$, $P = 0.27$) or inhibited (c; 1-way ANOVA, $F_{1.05, 3.16} = 1.14$, $P = 0.37$) around action initiation. **(d-f)** Activity and modulation across training of A2A⁺ early action-initiation excited neurons ($N = 42$ neurons/4 mice; average 16.8 neurons/mouse, s.e.m. = 5.61). Heat map (d), Z-scored activity (e), and area under the receiver operating characteristic curve (auROC) modulation index (f) of early action-initiation excited neurons around action initiation across training. **(g-h)** Cross-session correlation of the activity around action initiation of each early action-initiation excited neuron (g; 2-way ANCOVA, Neuron activity: $F_{1, 38} = 7.59$, $P = 0.009$; Training: $F_{1, 38} = 1.12$, $P = 0.30$; Activity x Training: $F_{1, 38} = 1.04$, $P = 0.32$) or the population activity of these neurons (h; 2-way ANOVA, Neuron activity: $F_{1, 3} =$

3.07, $P = 0.18$; Training: $F_{1, 3} = 0.68$, $P = 0.47$; Activity x Training: $F_{1, 3} = 0.64$, $P = 0.48$). **(i)** Behavior class decoding accuracy from A2A$^+$ early action-initiation excited neuron activity in the 0.5-s period prior to event compared to shuffled control. Line at 0.25 = chance. 3-way ANOVA, Neuron activity (v. shuffled): $F_{1, 3} = 169.03$, $P < 0.001$; Training session: $F_{2, 6} = 5.55$, $P = 0.04$; Behavior class: $F_{3, 9} = 50.84$, $P < 0.001$; Neuron activity x Training: $F_{2, 6} = 5.27$, $P = 0.05$; Neuron activity x Behavior: $F_{3, 9} = 4.63$, $P = 0.03$; Training x Behavior: $F_{6, 18} = 1.79$, $P = 0.16$; Neuron activity x Training x Behavior: $F_{6, 18} = 0.34$, $P = 0.91$. **(j-l)** Linear decoding accuracy from A2A$^+$ early action-initiation excited neuron activity of lever-press rate (j; 2-way ANOVA, Neuron activity: $F_{1, 3} = 3.76$, $P = 0.14$; Training: $F_{1.33, 3.99} = 4.49$, $P = 0.19$; Neuron activity x Training: $F_{1.32, 3.97} = 2.52$, $P = 0.82$), rate of switching from lever press to food-port entry or vice versa (k; 2-way ANOVA, Neuron activity: $F_{1, 3} = 60.87$, $P = 0.004$; Training: $F_{1.15, 3.45} = 2.84$, $P = 0.18$; Neuron activity x Training: $F_{1.38, 4.13} = 1.67$, $P = 0.28$), log-transformed current interpress interval (l; 2-way ANOVA, Neuron activity: $F_{1, 3} = 31.18$, $P = 0.01$; Training: $F_{1.62, 4.85} = 3.10$, $P = 0.14$; Neuron activity x Training: $F_{1.16, 4.48} = 3.03$, $P = 0.17$), and the prior press interpress interval (s) (m; 2-way ANOVA, Neuron activity: $F_{1, 3} = 2.95$, $P = 0.18$; Training: $F_{1.03, 3.09} = 1.73$, $P = 0.27$; Neuron activity x Training: $F_{1.04, 3.13} = 1.76$, $P = 0.28$). R = correlation coefficient between actual and decoded press rate. **(n)** Cross-session decoding accuracy of lever-press rate from the activity of A2A$^+$ early action-initiation-excited neuron population activity on the 1st training session. Planned, Bonferroni corrected, 2-tailed t-tests, Early: $t_6 = 2.94$, $P = 0.08$, 95% CI -0.03 - 0.55; Middle: $t_6 = 2.92$, $P = 0.08$, 95% CI -0.03 - 0.55; Overtrain: $t_6 = 0.37$, $P > 0.9999$, 95% CI -0.26 - 0.33. **(o-q)** Activity and modulation across training of A2A$^+$ overtrain action-initiation excited neurons ($N = 25$ neurons/4 mice; average 6.25 neurons/mouse, s.e.m. 2.69). Heat map (o), Z-scored activity (p), and auROC modulation index (q) of overtrain action-initiation excited neurons around action initiation across training. **(r-s)** Cross-session correlation of the activity around action initiation of each overtrain action-initiation excited neuron (r; 2-way ANCOVA, Neuron activity: $F_{1, 23} = 7.13$, $P = 0.01$; Training: $F_{1, 23} = 0.18$, $P = 0.68$; Activity x Time: $F_{1, 23} = 0.14$, $P = 0.72$) or the population activity of these neurons (s; 2-way ANOVA, Neuron activity: $F_{1, 3} = 2.01$, $P = 0.25$; Training: $F_{1, 3} = 0.20$, $P = 0.69$; Activity x Time: $F_{1, 3} = 0.31$, $P = 0.62$). **(t)** Behavior class decoding accuracy from A2A$^+$ overtrain action-initiation excited neuron activity. 3-way ANOVA, Neuron activity: $F_{1, 3} = 1296.08$, $P < 0.001$; Training session: $F_{2, 6} = 2.75$, $P = 0.14$; Behavior class: $F_{3, 9} = 32.57$, $P < 0.001$; Neuron activity x Training: $F_{2, 6} = 1.99$, $P = 0.22$; Neuron activity x Behavior: $F_{3, 9} = 11.86$, $P = 0.002$; Training x Behavior: $F_{6, 18} = 2.36$, $P = 0.07$; Neuron activity x Training x Behavior: $F_{6, 18} = 2.91$, $P = 0.04$. **(u-x)** Linear decoding accuracy from A2A$^+$ overtrain action-initiation excited neuron activity of lever-press rate (u; 2-way ANOVA, Neuron activity: $F_{1, 3} = 1.92$, $P = 0.26$; Training: $F_{1.66, 4.96} = 0.57$, $P = 0.57$; Neuron activity x Training: $F_{1.63, 4.88} = 0.54$, $P = 0.58$), switch rate (v; 2-way ANOVA, Neuron activity: $F_{1, 3} = 8.93$, $P = 0.06$; Training: $F_{1.23, 3.69} = 0.16$, $P = 0.76$; Neuron activity x Training: $F_{1.24, 3.71} = 0.17$, $P = 0.75$), log-transformed current interpress interval (w; 2-way ANOVA, Neuron activity: $F_{1, 3} = 8.56$, $P = 0.06$; Training: $F_{1.05, 3.15} = 0.92$, $P = 0.41$; Neuron activity x Training: $F_{1.08, 3.24} = 0.90$, $P = 0.42$), and the prior press interpress interval (s) (x; 2-way ANOVA, Neuron activity: $F_{1, 3} = 3.10$, $P = 0.18$; Training: $F_{1.76, 3.18} = 5.23$, $P = 0.28$; Neuron activity x Training: $F_{1.72, 5.17} = 1.52$, $P = 0.30$). **(y)** Cross-session decoding accuracy of lever-press rate from the activity of A2A$^+$ overtrain action-initiation-excited neuron population activity on the 8th training session. Planned, Bonferroni corrected, 2-tailed t-tests, Early: $t_6 = 0.61$, $P > 0.999$, 95% CI -0.27 to 0.39; Middle: $t_6 = 2.63$, $P = 0.12$, 95% CI -0.07 - 0.59; Overtrain: $t_6 = 2.86$, $P = 0.09$, 95% CI -0.04 - 0.62. **(z)** Percent of A2A$^+$ early action-initiation excited neurons that continued to be significantly excited by action initiation on the 4th and 8th training sessions. 2-tailed t-test, $t_3 = 0.37$, $P = 0.74$, 95% CI -53.97 - 42.86. **(aa)** Percent of A2A$^+$ overtrain action-initiation excited neurons that were also significantly excited by action initiation on 1st and 4th training sessions. 2-tailed t-test, $t_3 = 0.13$, $P = 0.31$, 95% CI -18.88 - 17.44. **(ab)** Percent of all coregistered D1$^+$ and A2A$^+$ neurons that were significantly modulated by action initiation across all phases of training ('stable action-initiation ensemble'). 2-tailed t-test, $t_3 = 3.55$, $P = 0.01$, 95% CI -7.34 - -1.35. **(ac)** Correlation between the stability of the early action-initiation excited A2A$^+$ ensemble (average auROC modulation index of this ensemble prior to action initiation during overtraining) and devaluation index for all subjects (A2A-cre: $N = 7$; 4 habitual, 3 goal-directed). Pearson $r_7 = -0.82$, $P = 0.02$. **(ad-ae)** Accuracy with which lever-press rate can be can be decoded from the activity of A2A$^+$ early (ab, 2-way ANOVA, Neuron activity: $F_{1, 3} = 7.44$, $P = 0.07$; Training: $F_{1.40, 4.21} = 0.17$, $P = 0.77$; Neuron activity x Training: $F_{1.41, 4.22} = 0.16$, $P = 0.79$) or overtrain (ae, 2-way ANOVA, Neuron activity: $F_{1, 3} = 30.03$, $P = 0.01$; Training: $F_{1.52, 4.57} = 0.45$, $P = 0.61$; Neuron activity x Training: $F_{1.51, 4.54} = 0.48$, $P = 0.60$) action-initiation inhibited neurons. A2A-cre: $N = 4$ (1 male). Data presented as mean ± s.e.m. Males = closed circles/solid lines, Females = open circles/dashed lines. *$P < 0.05$, **$P < 0.01$.

DMS A2A$^+$ neurons only transiently support action-outcome learning.

DMS A2A$^+$ neuronal activity is necessary for action-outcome learning. Given that a subpopulation of DMS A2A$^+$ neurons encodes actions and conveys information related to the extent of agency during early learning, we next reasoned that this neuronal activity might enable action-outcome learning to support goal-directed behavioral control. To test this, we chemogenetically inhibited DMS A2A$^+$ neurons during learning and then probed behavioral strategy using the devaluation test. We expressed hM4Di or a fluorophore control selectively in DMS A2A$^+$ neurons of A2A-cre or wildtype control mice (Figure 6a-b). Mice were trained to lever press to earn food-pellet rewards (Figure 6c). Prior to each of the 4 random-interval training sessions, mice received CNO (2.0 mg/kg, i.p.) to, in hM4Di-expressing A2A-cre subjects, inactivate A2A$^+$ neurons (Extended Data Figure 3-1). Chemogenetic inactivation of A2A$^+$ neurons did not alter acquisition of the instrumental lever-press behavior (Figure 6d; see also Extended Data Figure 6-1). Thus, DMS A2A$^+$ neurons are not necessary for general acquisition or execution of instrumental behavior. A2A$^+$ neuron inhibition did cause more entries into the food-delivery port, suggesting an inability to suppress switches to this competing behavior (Extended Data Figure 6-

1), but did not affect behavioral strategy variables (Extended Data Figure 6-2). Inhibition of A2A$^+$ neurons did disrupt the action-outcome learning underlying goal-directed decision making as evidenced by subsequent insensitivity to outcome devaluation (Figure 6e-f). Thus, DMS A2A$^+$ neurons are active during learning and this activity enables the action-outcome learning that supports agency for flexible, goal-directed decision making.

DMS A2A$^+$ neuronal activity is not sufficient to promote goal-directed behavioral control. We next asked whether DMS A2A$^+$ neuron activity drives agency. If so, augmenting A2A$^+$ neuronal activity should promote the dominance of goal-directed behavioral control after overtraining. However, we reasoned that it would not because habit formation is associated with reorganization of DMS A2A$^+$ neurons to an excited ensemble that does not significantly convey behavioral-strategy information. To test this, we chemogenetically activated A2A$^+$ neurons during learning and then probed behavioral strategy using the devaluation test. We expressed hM3Dq or a fluorophore control selectively in DMS A2A$^+$ neurons of A2A-cre or wildtype control mice (Figure 6g-h). Mice were trained to lever press to earn food-pellet rewards (Figure 6i) and were overtrained. Prior to each training session, mice received CNO (0.2 mg/kg, i.p.) to, in hM3Dq-expressing A2A-cre subjects, activate A2A$^+$ neurons (Extended Data Figure 3-1). Chemogenetic activation of A2A$^+$ neurons did not alter acquisition of the instrumental lever-press behavior (Figure 6j). It also did not prevent habit formation (Figure 6k-l). Both controls and subjects for which we activated A2A$^+$ neurons during learning were insensitive to devaluation at test. Thus, DMS A2A$^+$ neuronal activity is not sufficient to drive dominance of goal-directed behavioral control or disrupt habit formation.

DMS A2A$^+$ neuronal activity is not necessary to express goal-directed behavioral control. We next asked whether DMS A2A$^+$ neurons support the online use of agency for flexible, goal-directed decision making. We reasoned they would not because these neurons do not encode the action-outcome association needed for agency and only transiently encode actions. To test this, we chemogenetically inhibited DMS A2A$^+$ neurons at test, rather than during learning. We expressed hM4Di or a fluorophore control selectively in DMS A2A$^+$ neurons of A2A-cre or wildtype control mice (Figure 6m-n). Mice received limited training to lever press to earn food-pellet rewards (Figure 6o). All mice acquired the instrumental behavior (Figure 6p). Mice then received the outcome-specific devaluation test and were given CNO (2.0 mg/kg, i.p.) after sensory-specific satiety, prior to the lever-pressing probe test. Presses were expressed as a percent of baseline (training session prior to test) to normalize the pre-existing differences in press rate between groups (see Extended Data Figure 6-3 for press rate). Inactivation of A2A$^+$ neurons did not affect flexible, goal-directed decision making, as evidenced by sensitivity to devaluation in both groups (Figure 6q-r). Thus, A2A$^+$ neurons only transiently support action-outcome learning and become unnecessary for using such agency for flexible, goal-directed decision making.

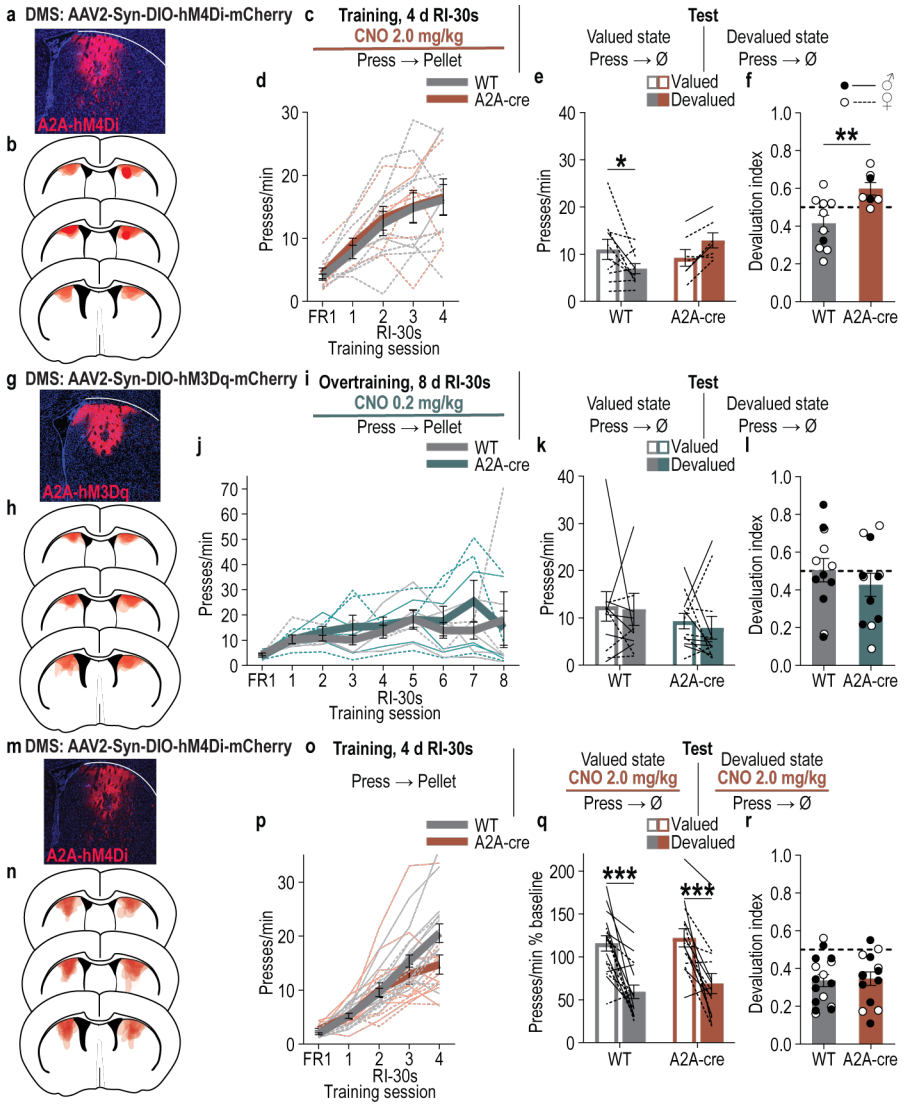


Figure 6: DMS A2A⁺ neurons are transiently necessary for action-outcome learning, but not goal-directed decision making. (a-f) Chemogenetic inactivation of DMS A2A⁺ neurons during learning. WT: N = 10 (1 male); A2A-cre: N = 7 (2 males). **(a)** Representative immunofluorescent image of cre-dependent hM4Di expression in DMS. **(b)** Map of DMS cre-dependent hM4Di expression for all subjects. **(c)** Procedure. RI, random-interval reinforcement schedule; CNO, clozapine-N-oxide. **(d)** Training press rate. 2-way ANOVA, Training: $F_{2.39, 35.84} = 30.54$, $P < 0.0001$; Genotype: $F_{1, 15} = 0.07$, $P = 0.79$; Training x Genotype: $F_{4, 60} = 0.05$, $P = 0.99$. **(e)** Test press rate. 2-way ANOVA, Value x Genotype: $F_{1, 15} = 9.67$, $P = 0.007$; Value: $F_{1, 15} = 0.02$, $P = 0.88$; Genotype: $F_{1, 15} = 0.96$, $P = 0.34$. **(f)** Devaluation index [(Devalued condition presses)/(Valued condition presses ⁺ Devalued presses)]. 2-tailed t-test, $t_{15} = 3.10$; $P = 0.007$, 95% CI 0.057 - 0.31. **(g-l)** Chemogenetic activation of A2A⁺ neurons during overtraining. WT: N = 12 (7 males); A2A-cre: N = 12 (6 males). **(g)** Representative immunofluorescent image of DMS cre-dependent hM3Dq expression. **(h)** Map of DMS cre-dependent hM3Dq expression. **(i)** Procedure. **(j)** Training press rate. 2-way ANOVA, Training: $F_{1.87, 41.22} = 15.50$, $P < 0.0001$; Genotype: $F_{1, 22} = 0.80$, $P = 0.38$; Training x Genotype: $F_{8, 176} = 0.25$, $P = 0.98$. **(k)** Test press rate. 2-way ANOVA, Value x Genotype: $F_{1, 22} = 0.03$, $P = 0.87$; Value: $F_{1, 22} = 0.20$, $P = 0.66$; Genotype: $F_{1, 22} = 1.30$, $P = 0.27$. **(l)** Devaluation index. 2-tailed t-test, $t_{22} = 0.89$; $P = 0.38$, 95% CI -0.26 - 0.10. **(m-r)** Chemogenetic inactivation of A2A⁺ neurons during test of behavioral strategy after learning. WT: N = 16 (9 males); A2A-cre: N = 14 (8 males). **(m)** Representative immunofluorescent image of cre-dependent hM4Di expression in DMS. **(n)** Map of DMS cre-dependent hM4Di expression. **(o)** Procedure. **(p)** Training press rate. 2-way ANOVA, Training: $F_{1.78, 49.82} = 107.5$, $P < 0.0001$; Genotype: $F_{1, 28} = 1.30$, $P = 0.26$; Training x Genotype: $F_{4, 112} = 5.08$, $P = 0.008$. **(q)** Test press rate normalized to pre-test training baseline. 2-way ANOVA, Value x Genotype: $F_{1, 28} = 0.04$, $P = 0.84$; Value: $F_{1, 28} = 47.56$, $P < 0.0001$; Genotype: $F_{1, 28} = 0.50$, $P = 0.49$. **(r)** Devaluation index. 2-tailed t-test, $t_{28} = 0.22$; $P = 0.83$, 95% CI -0.09 to 0.11. Data presented as mean ± s.e.m. Males = closed circles/solid lines, Females = open circles/dashed lines. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

DISCUSSION

Here we asked how the two major DMS neuronal output populations organize their activity to support action-outcome learning for flexible goal-directed decision making and how this changes with habit formation. The data reveal DMS cell-type specific encoding stability and reorganization. Early in learning both DMS neuron subtypes

encode actions and information underlying behavioral strategy to support acquisition of the action-outcome knowledge underlying agency. Whereas this action-outcome model is encoded and maintained in D1$^+$ neurons to enable stability of agency and model-based decision making, A2A$^+$ neurons reorganize their action encoding with habit formation.

DMS D1$^+$ neurons stably encode actions, convey engagement and organizational information underlying behavioral strategy, and integrate information about reward outcomes into the stable action representation. The encoding stability and convergence of action and outcome encoding suggests DMS D1$^+$ neurons form a lasting representation of learned contingencies that supports flexible, goal-directed decision making. This likely results from post-synaptic plasticity in these neurons following action-outcome learning[81]. Correspondingly, we found that D1$^+$ neurons have functional stability, supporting both action-outcome learning and the application of this learned agency for flexible, goal-directed decision making. Activating D1$^+$ neurons promotes the dominance of flexible, goal-directed decision making, similar to prior reports[82]. Together with evidence of stable action and action-outcome encoding, this indicates DMS D1$^+$ neurons maintain a representation of one's agency even once behavioral strategy has transitioned to a habit system independent of these neurons. This function in action-outcome learning and goal-directed decision making is likely achieved via the D1$^+$ direct pathway[83,84] that is a critical component of the basal ganglia circuit that drives motor execution[69,85]. Through learning, in the DMS, neurons in this pathway come to stably encode actions, behavioral strategy variables, and the action-outcome association to allow learned motor behaviors to be executed with adaptive agency.

DMS A2A$^+$ neurons transiently encode actions and behavioral transitions to support initial development of the action-outcome model. A subpopulation of A2A$^+$ neurons transiently encodes action initiation early in learning. Correspondingly, A2A$^+$ neurons support early action-outcome learning, but do so by conveying information distinct from D1$^+$ neurons. A2A$^+$ neurons convey behavioral transitions, but do not strongly convey other behavioral strategy variables. They also show little encoding of reward outcomes, consistent with prior reports[86], and do not represent the action-outcome association. That both D1$^+$ and A2A$^+$ neurons are active prior to instrumental actions and required for initial action-outcome learning is consistent with the coordinated function view[71,87,88] in which D1$^+$ neurons drive selected actions and D2/A2A$^+$ neurons permit such actions by inhibiting competing or unrewarded actions[68,69,71-73,87-93]. By inhibiting competing behaviors, DMS A2A$^+$ neurons could help establish appropriate action-outcome relationships, underlying their early necessity for such learning. Indeed, DMS A2A$^+$ neurons encode behavioral transitions and their inhibition caused more transitions to the competing food-port checking behavior. Thus, DMS D1$^+$ and A2A$^+$ neurons contribute unique information to together support the initial acquisition of the action-outcome knowledge needed for agency, extending the coordinated D1$^+$ and D2/A2A$^+$ function model beyond motor control to learning.

DMS A2A$^+$ neurons reorganize their action encoding with habit formation. With continued training the subpopulation of A2A$^+$ neurons that initially encodes actions ceases to do so. Instead, another ensemble of A2A$^+$ neurons gradually gains action encoding as habits form. This subpopulation conveys information about behavioral performance, but not strategy, indicating a shift in not only which neurons encode the action, but also the

information content they convey. Realignment was incomplete if habits did not form. This reorganization of action encoding may, therefore, contribute to a shift in the function of DMS A2A$^+$ neurons. In support of this, although A2A$^+$ neurons initially support action-outcome learning, they are unnecessary for the expression of such agency for goal-directed decision making and their activation is not sufficient to promote goal-directed behavior. Moreover, DMS A2A$^+$ neurons have been implicated in habit[94-96] and other behavioral strategy shifts including action-outcome reversal[83,97] and extinction[97]. Thus, the DMS A2A$^+$ neuron action subpopulation and coding properties reorganizes as the strategy generating those actions shifts.

The discoveries here open the door to many important future questions. The stability of D1$^+$ and reorganization of A2A$^+$ neuronal encoding likely reflect a combination of molecular[101-103], cellular[81], and circuit-level[50,103-105] mechanisms. These likely involve input from the prelimbic and orbitofrontal cortex[50,103,104,106], basolateral and central amygdala input[107], and dopamine modulation[108-110]. Future studies will be crucial to define these mechanisms. It will also be important to understand how striatal interneurons[111], astrocytes[112], and collateral interactions are involved[95,113,114]. Because DMS indirect striatopallidal projections are not necessary for action-outcome learning[83], the DMS A2A$^+$ neurons we find that function in initial action-outcome learning may be a distinct subtype[57,98-100], perhaps a non-canonical projection. This possibility warrants sfurther investigation coupling functional, molecular, and projection-target profiling. The task here was necessarily simple to allow neurons to be tracked from the initial learning through habit formation. Thus, the features of the action and outcome encoded in the DMS D1$^+$ neuronal action-outcome representation is an important future question. Being critical for goal-directed decision making, action value[115] and outcome identity[116] are both likely. An intriguing larger question is whether stable D1$^+$ neuronal encoding and A2A$^+$ encoding realignment is a principle of striatal contributions to learning and behavioral strategy.

Adaptive decision making requires understanding actions can produce particular consequences. Here we find such agency is learned through the activity of DMS D1$^+$ and A2A$^+$ neurons and then maintained in DMS D1$^+$ neurons to support adaptive, goal-directed decision making. The transition to automatic habits is associated with reorganization of A2A$^+$ neuronal encoding. These data provide neuronal circuit insights into how we learn, how we decide, and how our behavior shifts to inflexible automaticity. This helps understand how stress[106,107], exposure to drugs and alcohol[27,51,117], neurodevelopmental factors[91,118], and aging[119] can lead to the disrupted decision making and maladaptive habits.

AUTHOR CONTRIBUTIONS

MM and KMW conceptualized and designed the experiments, interpreted the data, and wrote the paper. MM executed the experiments and analyzed the data. AL, NP, JYG, MDM, WG, Alicia W, MS, Anna W, JSP assisted with experiments. BSH conducted the on-devaluation-test chemogenetic inactivation experiments. SPB conducted the behavioral pilot experiment. JRG conducted the electrophysiological validation experiments with support from SMH, CC, and MSL. NKG managed the breeding colonies and assisted with histological

verification. GJB, ACS, HTB, and AMW provided guidance and assistance with data analysis. PDB contribute guidance on behavioral modeling. AMW, HTB, and PDB provided guidance on data interpretation.

ACKNOWLEDGEMENTS

This research was supported by NIH R01DA046679 (KMW), NIH R01DA058374 (KMW), BBRF 32148 (KMW), NIH T32DA024635 (JRG), NIH F32DA056201 (JRG), A.P. Giannini Fellowship (JRG), NIH K99MH135177 (JRG), NIH TL4GM118977 (NP), NIH F32MH135680 (GJB), NSF NeuroNex 1707408 (HTB), and the Staglin Center for Behavior and Brain Sciences. The authors thank Dr. Sean Ostlund for helpful comments on the behavioral analysis.

COMPETING FINANCIAL INTERESTS

The authors have no biomedical financial interests or potential conflicts of interest to declare.

METHODS

Subjects

Male wildtype C57/Bl6J mice (Jackson Laboratories, Bar Harbor, ME) were used for the behavioral pilot experiment. All other experiments were conducted with male and female Drd1a-Cre[120] (D1-cre) and Adora2A-Cre[121] (A2A-cre) transgenic mice and their wildtype littermates bred in house. Mice were between 9 - 16 weeks old at the time of experiment onset/surgery. Mice were housed in a temperature (68 - 79 °F) and humidity (30 - 70%) regulated vivarium on a 12:12 hr reverse dark/light cycle (lights off at 7 AM). Mice were initially housed in same-sex groups of 3 - 4 mice/cage. Prior to experiment onset/surgery mice were single-housed to facilitate food deprivation and preserve implants. Unless noted below, mice were provided with food (standard rodent chow, Lab Diet, St. Louis, MO) and water *ad libitum* in the home cage. Mice were handled for 3 - 5 days prior to the start of behavioral training for each experiment. All procedures were conducted in accordance with the NIH Guide for the Care and Use of Laboratory Animals and were approved by the UCLA Institutional Animal Care and Use Committee.

Surgery

Mice were anesthetized with isoflurane (3% induction, 1% maintenance), and positioned in a digital stereotaxic frame (Kopf, Tujunga, CA). Subcutaneous Rimadyl (Carprofen; 5 mg/kg; Zoetis, Parsippany, NJ) was given pre-operatively for analgesia and anti-inflammatory purposes. An incision was made along the midline to expose the skull. For viral infusions, after performing a small craniotomy, virus was injected using a 28-g infusion needle (PlasticsOne, Roanoke, VA) connected to a 1-mL syringe (Hamilton Company, Reno, NV) by intramedic polyethylene tubing (BD; Franklin Lakes, NJ) and controlled by a syringe pump (Harvard Apparatus, Holliston, MA). Virus was injected at a rate of 0.1 µl/min and the needle was left in place for 10 min post-injection. Further experiment-specific surgical details are provided below. After surgery, mice were kept on a heating pad maintained at 35 °C for 1 hr and then single-housed in a clean homecage for recovery and monitoring. Mice received chow containing the antibiotic TMS for 7 days following surgery to prevent infection, after which they were returned to standard rodent chow.

Behavioral procedures

Apparatus. Training took place in Med Associates (East Fairfield, VT) wide mouse operant chambers housed within sound- and light-attenuating boxes. Each chamber had metal grid floors and contained a retractable lever to the left of a recessed food-delivery port (magazine) on the front wall. A photobeam entry detector was positioned at the entry to the food port. Each chamber was equipped with 2 pellet dispensers to deliver either 20-mg grain or chocolate-flavored purified pellets (Bio-Serv, Frenchtown, NJ) into the food port when activated. A fan mounted to the outer chamber provided ventilation and external noise reduction. A 3-watt, 24-volt house light mounted on the top of the back wall opposite the food port provided illumination. To monitor subject behavior, monochrome digital cameras (Med Associates) were positioned with a top-down view of the conditioning chambers.

Food deprivation. 3 - 5 days prior to the start of behavioral training, mice were food-deprived to maintain 85% - 90% of their free-feeding body weight. Mice were given 1.5 - 3.0 g of their home chow at the same time daily at least 2 hrs after training.

Outcome pre-exposure. To familiarize subjects with the food pellet that would become the instrumental outcome, mice were given 1 session of outcome pre-exposure. Mice were placed in a clean, empty cage and allowed to consume 30 of the food pellets from a metal cup. If any pellets remained, they were placed in the home cage overnight for consumption.

Magazine conditioning. Mice received 1 session of training in the operant chamber to learn where to receive the food pellets (20-mg grain or chocolate-purified pellets). Mice received 30 non-contingent pellet deliveries from the food port with a fixed 60-s intertrial interval.

Instrumental Training. Mice next received 1 session/day consecutively of instrumental conditioning in which lever presses earned delivery of a single food pellet. The same experimenter conducted each session of training and subsequent tests. Earned pellet type (grain or chocolate) was counterbalanced across subjects within each group of each experiment. Each session began with the illumination of the house light and insertion of the lever and ended with the retraction of the lever and turning off of the house light. Each training session ended after 30 outcomes had been earned or 30 min elapsed. In all cases, instrumental training began on a fixed-ratio 1 schedule (FR-1), in which each action was reinforced with one food pellet outcome. Once 90% of the maximum

session outcomes were earned, the reinforcement schedule was shifted to random-interval (RI) in which a variable interval must elapse following a reinforcer for another press to be reinforced. Mice received either 4 or 8 (overtrained) consecutive sessions of RI training starting first with 1 session on a RI-15s schedule, and then the remaining sessions on a RI-30s schedule.

Alternate outcome exposures. To equate exposure of the non-trained pellet, all mice were given non-contingent access to the same number of the alternate food pellets as the earned pellet type (e.g., chocolate pellets if grain pellets served as the training outcome) in a different context (clear plexiglass cage).

Sensory-specific satiety outcome devaluation test. Testing began 24 hr after the final instrumental conditioning session. Mice were given 1 - 1.5 hr access to either 4 g of the food pellets previously earned by lever pressing (Devalued condition) or 4 g of the non-trained pellets to control for general satiety (Valued condition). The remaining pellets were weighed following prefeeding to measure total consumption. Consumption did not significantly differ between the Devalued v. Valued conditions or between the control and experimental group for any experiment (Supplemental Table 2). Immediately after this prefeeding, lever pressing was assessed during a brief, 5-min, non-reinforced probe test. To assess the efficacy of the sensory-specific satiety, following the probe test, mice were given a 10-min consumption choice test with simultaneous access to 1 g of both pellet types. To ensure that lever-press test data was not confounded by incomplete devaluation, subjects that failed to fully reject the devalued food type (i.e., consumed more than our margin of measurement error, 0.09 g, of the pre-fed pellet type during post-test consumption) were not included in the analysis (see exclusions noted for each experiment below). The remaining mice consumed less of the prefed pellet than non-prefed pellet, indicating successful sensory-specific satiety devaluation (Supplemental Table 3). 24 - 48 hr after the first devaluation test, mice received 1 session of instrumental retraining (RI-30s), followed the next day by a second devaluation test in which they were prefed the opposite food pellet. Thus, each mouse was tested in both the Valued and Devalued conditions, with test order counterbalanced across subjects within each group for each experiment.

Establishing goal-directed behavioral strategy following limited instrumental random-interval training and habit following overtraining

Naïve, male C57BL/6J mice (8 weeks old, Jackson Laboratory) were used to establish the training protocols for action-outcome learning and goal-directed decision making (N = 8) or habit formation with overtraining (N = 8). No subjects were excluded. Mice were food restricted and then began instrumental training, as described above. Mice were habituated to intraperitoneal (i.p.) injections during the final day of FR-1 training. Then, all subjects received an i.p. injection of 0.9% saline following each of the RI training sessions. Following training, mice received a counterbalanced pair of sensory-specific satiety outcome-specific devaluation tests, as above. No injection was given on test.

Microendoscopic calcium imaging

Naïve, male and female Drd1a-cre and A2A-cre mice (D1-cre: Final N = 4, 1 male; A2A-cre habitual N = 4, 1 male; A2A-cre goal-directed N = 4, 3 male) were used in this experiment to monitor calcium activity in individual $D1^+$ or $A2A^+$ DMS striatal projection neurons during instrumental conditioning. 18 (D1-cre: 9, A2A-cre: 9) subjects with lack of viral expression or without detectable individual neurons were excluded from the experiment. 3 (D1-cre: 2, A2A-cre: 1) subjects with viral overexpression were excluded from the dataset. At surgery, mice were infused with an adeno-associated virus (AAV) encoding the cre-dependent genetically encoded calcium indicator jGCaMP7s (0.40 µl; AAV9-syn-Flex-jGCaMP7s) into the DMS (AP +0.15, ML ±1.9, DV −2.65 mm from bregma). In a separate surgery, 4 - 7 days later, a unilateral 1.1-mm craniotomy (left/right hemisphere counterbalanced across subjects) was performed, centered above the DMS (AP +0.20, ML ±1.70 mm from bregma). Cortex was aspirated with a 27 - 30-g needle to a depth of approximately 1.8 mm. A 1-mm diameter, 4-mm long gradient refractive index (GRIN) lens (Inscopix, Palo Alto, CA) was implanted above the DMS (DV -2.25 mm). The lens was fixed to the skull with cyanoacrylate glue and secured with C&B Metabond quick adhesive cement system (Parkell Inc., Edgewood, NY), followed by opaque dental cement (Lang Dental Manufacturing, Wheeling, IL). Approximately 3 - 4 weeks following implant, a baseplate was attached to the previously formed headcap, ensuring proper alignment between the GRIN lens and the miniscope.

Mice were habituated to restraint and fitted with the miniscope for 2 days prior to behavioral training. Mice received a second magazine training session to become accustomed to retrieving pellets with the head-mounted miniscope. After the first FR-1 instrumental session, mice were subsequently fitted with the miniscope prior to each of 8 RI training sessions. Following training, mice received a pair of sensory-specific satiety outcome-

specific devaluation tests, as above. We noticed that not all of the A2A-cre subjects developed habits (insensitivity to devaluation) with overtraining. We, therefore, used a mean split to separate these subjects into those that formed habits (Devaluation index ≥ 0.483; N = 4) and those that remained sensitive to devaluation showing goal-directed decision making (Devaluation index < 0.483; N = 4). These groups did not differ in their acquisition of lever pressing during training (Training session: $F_{1.51, 9.04} = 8.89$, $P = 0.01$; Group: $F_{1, 6} = 0.48$, $P = 0.52$; Training x Group: $F_{8, 48} = 1.02$, $P = 0.43$). They also did not differ in the efficacy of devaluation (see Supplemental Tables 2 - 3).

Predicting behavior strategy. To reveal behavioral variables that relate to the extent of goal-directed v. habitual behavioral strategy we used a linear regression model to ask whether and how behavior during the final training session predicts subsequent sensitivity to outcome devaluation. We pooled behavioral data from all the D1-cre and A2A-cre subjects (N = 12) so we could exploit the variability in sensitivity to devaluation for this model. We included measures of behavioral engagement: lever-press rate and switch rate (rate at which subject switch from pressing to checking the food port or vice versa; a switch was defined as any time the animal switched from pressing to enter the food port or from enter the food port to press). We also included measures of behavioral patterning. On the random-interval reinforcement schedule a variable amount of time must elapse after an earned reward before another press will earn a reward. Thus, we focused our pattern measures of the time subjects waited in between presses (interpress interval, IPI). As a measure of overall pattern, we included the variability/consistency (standard deviation) of the log-transformed interpress intervals. As a measure of short-term patterning and the extent to which current trial behavior relates to past trial behavior, we included the correlation between the current interpress interval, n, and the interpress interval of the prior press, n-1[63,64]. In preliminary investigations, we also evaluated whether the influence of the interpress interval variables on devaluation sensitivity depended on whether there was a food-port check during the interval and found it did not and so focused on interpress interval. Using the MATLAB function *fitlm*, we built a linear regression model to measure the extent to which press rate, switch rate, standard deviation of the log interpress intervals, and n:n-1 interpress interval correlation predict subsequent sensitivity to devaluation, measured by the devaluation index. The model included the standard deviation of log-transformed interpress-intervals and n:n-1 interpress interval correlation as fixed effects, along with an interaction term between press rate and switch rate. Model performance was evaluated using MATLAB function *predict* in a 5-fold cross-validation procedure repeated 5 times. The average adjusted R-squared was calculated for each cross-validation and averaged across repetitions to provide a robust estimate of the model's predictive power.

Miniscope system. Calcium imaging videos were recorded using an open-source, head-mounted UCLA Miniscope imaging device, which interfaces with the UCLA open-source Miniscope DAQ hardware and software (http://miniscope.org/) to capture the neuronal calcium dynamics with corresponding frame timestamps. A webcam (Logitech, San Jose, CA) recorded the Med Associates interface LEDs, which display light pulses coincident with task events (lever presses, food-port entry, pellet delivery). This video was simultaneously saved on the same computer alongside the calcium imaging video via the miniscope software with synchronized frame timestamps.

Calcium trace extraction and spike inference. Videos were initially processed using custom Python code[122] to concatenate all the videos from a session into one tiff stack, spatially downsample by a factor of 2, and temporally downsample to an effective frame rate of ~7.5 frames/s to accelerate image processing. The Python implementation of the Calcium Imaging Analysis package, 'CaImAn'[123], was then used to process the videos. We applied the nonrigid motion correction algorithm to remove tissue motion artifacts and then applied constrained non-negative matrix factorization for microendoscopic data to extract individual spatial contours and demix temporal fluorescence traces for each detected neuron. Individual fluorescent signals were deconvolved to estimate discrete spikes for each neuron[124]. The deconvolved traces, the estimation of discrete spikes that generate the calcium traces, were derived from the denoised fluorescence traces using CaImAn's '*deconvolveCa*' function. Neurons were coregistered across sessions using the open source package CellReg[125]. Activity for each neuron was Z-scored to the whole session to allow comparison of baseline and phasic activity across cells and subjects with different levels of GCaMP expression and/or lens proximity.

Behavior extraction. All behavior-event videos from a session were concatenated and downsampled as above. Using custom MATLAB code, the brightness of each LED on the interface was measured for each frame. A task event was considered to have occurred if the LED brightness was >2x the background brightness threshold (determined individually per recording), thus binarizing task events in each frame of the video. The extracted frame-by-frame event information was then interpolated to the timestamps of the calcium imaging video.

Cell classification. We used an analysis based on the area under the receiver operating characteristic curve to identify neurons tuned to key behavioral events: 'action initiation', first press in session, first press after earned reward, and first press after a non-reinforced food-port check; 'action termination', last press before earned reward collection and last press before a non-reinforced food-port check; 'food-port check', entries into the food delivery port, for those in which there was no reward present (a.k.a. non-reinforced entries); 'reward consumption', consumption of the earned food pellet occurring at the time of exit from the food-delivery port. The auROC provides a non-parametric measure of the separability between two distributions of data. To increase the temporal precision in linking neural activity to behavior, we used the deconvolved traces to distinguish between a neuronal activity preceding v. in response to an event. For each neuron, we defined two distributions of activity: one from the activity aligned to the ±2.9 s window around repeated occurrences of a specific behavioral event and one from equated baseline periods in which subjects were not engaged in the task. We compared these two distributions of neuronal activity using the MATLAB *perfcurve* function to generate receiver operating characteristic curves of the true positive rate and the false positive rate and measured the area under the curve (auROC). To determine if a neuron significantly encoded an event, we generated a null distribution of auROC values by circularly shifting the activity trace for each neuron and recalculating the auROC 1000 times. Neurons with true auROC values >97.5% of the auROCs in the null distribution were considered significantly excited by an event and those with true auROC values <2.5% of the auROCs in the null distribution were considered significantly suppressed. To resolve the time at which neurons were significantly modulated within the ±2.9 s window, the auROC was further calculated at each individual frame within the ±2.9 s window around each event.

Behavior decoding. We used support vector machines (SVM) to decode key behavioral events (action initiation, action termination, food-port check, and reward consumption) using the activity of specific neuron populations. We used the temporal fluorescence trace because it is a continuous signal that can reflect recent neural activity that contains a more integrated signal important for changes between behaviors. To predict behavior events prior to onset, observation data used to train and test the classifier consisted of the 0.4-s epoch preceding each of the 4 distinct event labels. Since there were large variations in the number of each type of behavioral event, we controlled for the number of each event label. First, for the behavior event labels that were overrepresented (usually food-port entries), we found the median number for all event labels, and limited those event labels to that value. Then, for the behavior event labels that were underrepresented, we used the synthetic minority oversampling technique (SMOTE)[126] to generate synthetic activity samples to match the event label with the highest number of events. For each session, the classifier was trained (MATLAB function *fitcecoc*) on ¾ of the data, randomly selected, to generate a model and behavior was predicted (MATLAB function *predict*) with the generated model on the remaining data. Decoding performance was averaged across 20 different random data splits. Decoding performance was assessed by computing the F-score, a standard measure of binary classification accuracy that weighs recall (proportion of actual positives that are correctly predicted) against precision (the proportion of positive predictions that are actually positive) of the classifier. It is calculated as the harmonic mean of precision and recall. We performed a 1000-fold shuffling procedure to confirm that the average decoding of randomized data is at the theoretical chance level of ¼ correct.

To examine whether the ongoing lever-press rate, switch rate, and interpress-interval fluctuation could be explained by the neuronal activity, we fit a linear model (MATLAB function: *fitlm*) using the temporal fluorescence traces as predictors and behavior rate as the response. The discrete lever presses were transformed into an ongoing behavior rate by convolving the lever presses with a half-gaussian-post event kernel width of 60 s and a standard deviation of the distribution of 6 (see Extended Data Figure 1-5a for example). Switch points were identified as a shift from a lever press to a food-port entry vice versa and were convolved with a gaussian kernel width of 60 s and standard deviation of 6. As a measure of how long the subject waited before pressing, relative to overall session press wait times, we used log interpress intervals. Interpress intervals were assigned to the frame in which the current press occurred and were defined as current press (n) time ─ the time of the last press (n-1) before being log transformed. These were also convolved with a half-gaussian-post event kernel width of 60 s and a standard deviation of 6. As a measure of past trial action performance, we also calculated the n-1 interpress intervals. These were assigned to the frame in which the current press (n) occurred and were defined as the time of the last press (n-1) - the time of the press before that (n-2). These were also convolved with a gaussian-post event kernel width of 60 s and a standard deviation of 6. The window of 60 s and standard deviation of 6 was used because of the slow decay time constant of GCaMP7s [127]. The predictor and response data were Z-scored. A model was generated using the initial ¾ of the data and behavior rates were predicted (MATLAB function *predict*) with the generated model on the remaining data. Decoding performance was

assessed by calculating the correlation between the predicted behavior rate and the actual behavior rate. We performed a 1000-fold circular shuffling procedure to compare decoding accuracy to that of randomized data.

Modulation fidelity. To determine whether neural activity in response to a behavior was consistent across sessions at the neuron level, we averaged the deconvolved activity of a single neuron across all trials within a session to get a single response trace for a session, and compared it to its average trace for each session using a Spearman correlation. For the whole population, we averaged the activity of all of the neurons to get a population-level average response trace for each session and correlated these population-level traces across sessions. To determine whether individual neuron activity was consistent within each session, we calculated the Spearman correlation for the activity of a single neuron between all trials and averaged across correlations to get an overall within-session reliability measure for each neuron. We assessed the whole population within-session reliability by calculating the average activity trace for all neurons for each individual trial, calculated the Spearman correlation between trial activity, and calculated the average correlation as a measure of the whole population within-session reliability. In all cases, we performed a 1000-fold circular shuffling procedure to compare modulation fidelity to that of randomized data. To further assess the reliability of the neural population on a trial-by-trial basis, we quantified the percentage of trials in which each neuron responded, using a response threshold of 2x standard deviation of the trace activity during the 2 s prior to behavior event onset. The distribution of percent of trial response for all neurons was plotted to assess the response heterogeneity of the neural population across sessions.

High-dimensional population vector analysis. Neural activity from individual neurons was aligned to two key behavioral events: action initiation and reward. For each event, temporal fluorescence traces were averaged across trials. We analyzed the temporal dynamics of the population activity using a sliding window to create a time series of population vectors for each event. The neural activity within a 4-frame (~ 0.5-s) time window was averaged to form a population vector representing the population's activity at a specific time point. This was repeated with a 1-frame timestep size to generate population vectors for each event over the entire ±2.9-s window around the events. The similarity between the population representations of action initiation and reward at each timepoint was quantified by calculating the cosine distance between the two population vectors (MATLAB function *pdist*) at each time point. The cosine distance is defined as $1-\cos(\theta)$, where θ is the angle between the two vectors. This metric was chosen because it is sensitive to the pattern of neural activity and is insensitive to the overall firing rate or vector magnitude. The distance was calculated for each sliding window, resulting in a trajectory of real distance values over time. The average and peak cosine distance were quantified across the ±2.5-s window around the events. To visualize representative examples, representative vector distances were selected at the timepoint where the average distance was greatest in the first training session. The magnitude of each vector was calculated using the Euclidean norm[65] as the square root of the sum of the squared activity of all neurons. The angle in degrees between the two vectors was calculated as the inverse cosine of 1-cosine distance.

Representational Similarity Analysis. We used a time-resolved Representational Similarity analysis to examine the geometry of the neural activity within the entire neural population for each occurrence of action initiation and reward across training session. For each session, an empirical Representational Dissimilarity Matrix (RDM) was constructed from the neural activity for all of the Action initiation and Reward occurrences. Each entry in the RDM represents the pairwise dissimilarity between the activity patterns of two events, calculated as 1 - Pearson's correlation coefficient between the population activity for each event. The resulting matrix is symmetric, with a diagonal of zeros. A theoretical model RDM was generated to hypothesize the representational geometry between the same event types (e.g., Action initiation-Action initiation or Reward-Reward) as highly similar, represented as a pairwise dissimilarity of 0, while the representational geometry between different event types (e.g., Action initiation-Reward) as distinct, represented as a pairwise dissimilarity of 1. We then used Spearman's rank correlation to compare the lower-triangular portion of the empirical RDM with the model RDM, using a sliding window, for each session. A high correlation between the empirical RDM and model RDM indicates the neural representation between action initiation and reward is dissimilar. A lower correlation indicates similar neural representations between events. A visualization of the empirical RDM's across sessions for a representative subject were generated using multidimensional scaling (MDS; MATLAB function *mdscale*), to apply dimensionality reduction to the matrix of pairwise dissimilarities, capturing the relative dissimilarities between all occurrences of action initiation and reward. The representative RDMs were selected at the timepoint where the correlation between the empirical and model RDMs had the highest correlation (a.k.a. highest dissimilarity). To determine the correlation expected by chance, we performed a 1000-fold permutation of the event labels and

recalculated the correlation. The observed correlation was considered significant if it exceeded the 95th percentile of the null distribution.

Chemogenetic inhibition of DMS D1⁺ neurons during instrumental learning

Naïve, male and female D1-cre mice (Final *N* = 9, 4 males) and wildtype littermate controls (*N* = 7, 2 males) served as subjects in this experiment to assess the necessity of DMS D1⁺ neuron activity for the action-outcome learning that supports goal-directed decision making. 3 D1-cre subjects with off-target viral expression were excluded from the dataset. 3 (D1-cre: 2, WT: 1) subjects were excluded for ineffective sensory-specific satiety. At surgery, all mice received bilateral infusion of AAV encoding a cre-inducible inhibitory designer receptor human M4 muscarinic receptor (hM4Di; AAV2-Syn-DIO-hM4Di-mCherry; 0.3 µl) into the DMS (AP +0.2, ML ±1.8, DV −2.65 mm from bregma). After 2 weeks of recovery, mice were food restricted and then began instrumental training, as described above. Mice were habituated to i.p. injections during the final day of FR-1 training. Then, all subjects received an i.p. injection of water-soluble clozapine-n-oxide (CNO; 2.0 mg/kg; Hello Bio, Princeton, NJ) 30 min prior to each of the 4 RI training sessions. Following training, mice received a pair of sensory-specific satiety outcome-specific devaluation tests, as above. No CNO was given on test. CNO was given prior to the retraining session (RI-30s) in between tests.

Chemogenetic activation of DMS D1⁺ neurons during instrumental overtraining

Naïve, male and female D1-cre mice (Final *N* = 9, 4 males) and wildtype littermate controls (*N* = 8, 5 males) served as subjects in this experiment to assess whether DMS D1⁺ neuron activity is sufficient to promote action-outcome learning for goal-directed behavioral strategy and prevent the habit formation that normally occurs with overtraining. 6 D1-cre subjects with off-target viral expression were excluded from the dataset. 4 (D1-cre: 1, WT: 3) subjects were excluded for ineffective sensory-specific satiety. 1 WT subject that ate twice as much prior to one test than the other was excluded from the dataset to prevent the confound of differential general satiety. At surgery, all mice received bilateral infusion of AAV encoding a cre-inducible excitatory designer receptor human M3 muscarinic receptor (hM3Dq; AAV2-Syn-DIO-hM3Dq-mCherry; 0.3 µl) into the DMS (AP +0.2, ML ±1.8, DV −2.65 mm from bregma). After 2 weeks of recovery, mice were food restricted and then began instrumental training, as described above. Mice were habituated to i.p. injections during the final day of FR-1 training. Then, all subjects received CNO (0.2 mg/kg i.p.; Hello Bio, Princeton, NJ) 30 min prior to each of 8 RI training sessions. Following training, mice received a pair of sensory-specific satiety outcome-specific devaluation tests, as above. No CNO was given on test days. CNO was given prior to the retraining session (RI-30s) in between tests.

Chemogenetic inhibition of DMS D1⁺ neurons during goal-directed decision making

Naïve, male and female D1-cre mice (Final *N* = 8, 2 males) and wildtype littermate controls (*N* = 12, 7 males) served as subjects in this experiment to assess whether DMS D1⁺ neuron activity is necessary for the expression of goal-directed decision making. 7 D1-cre subjects with off-target viral expression were excluded from the dataset. At surgery, all mice received bilateral infusion of AAV encoding cre-inducible hM4Di (AAV2-Syn-DIO-hM4Di-mCherry; 0.3 µl) into the DMS. After 2 weeks of recovery, mice were food restricted and then began instrumental training, as described above. Mice were habituated to i.p. injections during the final day of FR-1 training. Mice received an i.p. injection of 0.9% saline 30 min prior to each of 4 RI training sessions. Following training, mice received a pair of sensory-specific satiety outcome-specific devaluation tests, as above. Immediately after the pre-feeding, all mice received CNO (2.0 mg/kg i.p.) and commenced to the non-rewarded probe test 30 min later. Saline was given prior to the retraining session (RI-30s) in between tests.

Chemogenetic inhibition of DMS A2A⁺ neurons during instrumental learning

Naïve, male and female A2A-cre mice (Final *N* = 7, 2 males) and wildtype littermate controls (*N* = 10, 1 male) served as subjects in this experiment to assess whether DMS A2A⁺ neuron activity is necessary for the action-outcome learning that supports goal-directed decision making. 3 A2A-cre subjects with off-target viral expression were excluded from the dataset. 3 (A2A-cre: 1, WT: 2) subjects were excluded for ineffective sensory-specific satiety. At surgery, all mice received bilateral infusion of AAV encoding cre-inducible hM4Di (AAV2-Syn-DIO-hM4Di-mCherry; 0.3 µl) into the DMS. After 2 weeks of recovery, mice were food restricted and then began instrumental training, as described above. Mice were habituated to i.p. injections during the final day of FR-1 training. Then, all subjects received CNO (2.0 mg/kg i.p.) 30 min prior to each of 4 RI training sessions. Following training, mice received a pair of sensory-specific satiety outcome-specific devaluation tests, as above. No CNO was given on test. CNO was given prior to the retraining session (RI-30s) in between tests.

Chemogenetic activation of DMS A2A$^+$ neurons during instrumental overtraining

Naïve, male and female A2A-cre mice (Final *N* = 12, 6 males) and wildtype littermate controls (*N* = 12, 7 males) served as subjects in this experiment to assess whether DMS A2A$^+$ neuron activity is sufficient to promote action-outcome learning for goal-directed behavioral strategy and prevent the habit formation that normally occurs with overtraining. 8 A2A-cre subjects with off-target viral expression were excluded from the dataset. 1 WT subject was excluded due to mechanical lesion. 11 (A2A: 6, WT: 5) subjects were excluded for ineffective sensory-specific satiety. At surgery, all mice received bilateral infusion of AAV encoding a cre-inducible hM3Dq (AAV2-Syn-DIO-hM3Dq-mCherry; 0.3 µl) into the DMS. After 2 weeks of recovery, mice were food restricted at then began instrumental training, as described above. Mice were habituated to i.p. injections during the final day of FR-1 training. Then, all subjects received CNO (0.2 mg/kg i.p.) 30 min prior to each of 8 RI training sessions. Following training, mice received a pair of sensory-specific satiety outcome-specific devaluation tests, as above. No CNO was given on test. CNO was given prior to the retraining session (RI-30s) in between tests.

Chemogenetic inhibition of DMS A2A$^+$ neurons during goal-directed decision making

Naïve, male and female A2A-cre mice (Final *N* = 14, 8 males) and wildtype littermate controls (*N* = 16, 7 males) served as subjects in this experiment to assess whether DMS A2A$^+$ neuron activity is necessary for the expression of goal-directed decision making. 6 A2A-cre subjects with off-target viral expression were excluded from the dataset. At surgery, all mice received bilateral infusion of AAV encoding cre-inducible hM4Di (AAV2-Syn-DIO-hM4Di-mCherry; 0.3 µl) into the DMS. After 2 weeks of recovery, mice were food restricted at then began instrumental training, as described above. Mice were habituated to i.p. injections during the final day of FR-1 training. Then, all subjects received saline (0.9% i.p.) 30 min prior to each of 4 RI training sessions. Following training, mice received a pair of sensory-specific satiety outcome-specific devaluation tests, as above. Immediately after the pre-feeding, all mice received CNO (2.0 mg/kg i.p.) and commenced to the non-rewarded probe test 30 min later. Saline was given prior to the retraining session (RI-30s) in between tests. A2A-cre and control mice had slightly different press rates on the last day of training, and so test data were normalized to baseline press rates.

Electrophysiological validation of chemogenetic manipulations

Whole-cell patch clamp recordings were used to validate the efficacy of chemogenetic manipulation of DMS D1$^+$ and A2A$^+$ neurons. Male and female naïve Drd1a-cre and A2A-cre mice served as subjects. At surgery, mice received bilateral infusion of AAV encoding cre-inducible hM4Di (AAV2-Syn-DIO-hM4Di-mCherry; 0.3 µl), cre-inducible hM3Dq (AAV2-Syn-DIO-hM3Dq-mCherry; 0.3 µl), or a fluorophore control (AAV2-Syn-DIO-mCherry; 0.3 µl) into the DMS. Recordings were performed from slices containing the DMS ~2 - 3 weeks following infusion. Mice were ~2 months old at the time slices were taken. Mice were deeply anesthetized with isoflurane and transcardially perfused with ice-cold, oxygenated NMDG-based slicing solution containing (in mM): 30 $NaHCO_3$, 20 HEPES, 1.25 NaH_2PO_4, 102 NMDG, 40 glucose, 3 KCl, 0.5 $CaCl2\text{-}2H_2O$, 10 $MgSO_4\text{-}H_2O$ (pH adjusted to 7.3-7.35, osmolality 300-310 mOsm/L). Brains were extracted and placed in ice-cold, oxygenated NMDG slicing solution. Coronal slices (300 µm) were cut using a vibrating microtome (VT1000S; Leica Microsystems, Germany), transferred to an incubating chamber containing oxygenated NMDG slicing solution warmed to 32-34 °C, and allowed to recover for 15 min before being transferred to an artificial cerebral spinal fluid (aCSF) solution containing (in mM): 130 NaCl, 3 KCl, 1.25 NaH_2PO_4, 26 $NaHCO_3$, 2 $MgCl_2$, 2 $CaCl_2$, and 10 glucose) oxygenated with 95% O_2, 5% CO_2 (pH 7.2-7.4, osmolality 290-310 mOsm/L, 32-34°C). After 15 min, slices were moved to room temperature and allowed to recover for ~30 additional min prior to recording. All recordings were performed using an upright microscope (Olympus BX51WI, Center Valley, PA) equipped with differential interference contrast optics and fluorescence imaging (QIACAM fast 1394 monochromatic camera 765 with Q-Capture Pro software, QImaging, Surrey, BC, Canada). Patch pipettes (3-5 MΩ resistance) contained a K-gluconate-based solution containing (in mM): 112.5 K-gluconate, 4 NaCl, 17.5 KCl, 0.5 $CaCl_2$, 1 $MgCl_2$, 5 ATP (K^+ salt), 1 NaGTP, 5 EGTA, 10 HEPES, pH 7.2 (270–280 mOsm/L). Biocytin (0.2%, Sigma-Aldrich, St. Louis, MO) was included in the internal recording solution for subsequent verification of fluorescence labeling in recorded cells. Recordings were obtained using a MultiClamp 700B Amplifier (Molecular Devices, Sunnyvale, CA) and the pCLAMP 10.7 acquisition software. Fluorescence-guided whole-cell patch clamp recordings in current-clamp mode were obtained from DMS D1$^+$ and A2A$^+$ neurons (Control, *N* = 19 cells, 4 D1-cre mice, 1 male; D1-cre, hm4Di: *N* = 8 cells, 4 mice, 3 males, hM3Dq: *N* = 10 cells, 5 mice, 3 males; A2A-cre, hm4Di: *N* = 8 cells, 4 mice, 3 males, hM3Dq: *N* = 10 cells, 5 mice, 3 males). After breaking through the membrane,

recordings were obtained from cells while injecting suprathreshold depolarizing current (1 s). Current injection intensities between 100 - 800 pA, resulting in 5 - 15 action potentials, were selected for recordings. Electrode access resistances were maintained at <30 MΩ throughout recordings. Number of action potentials generated were recorded both prior to and after bath application of CNO (10 or 100 μM) for >10 minutes. Percent change in the number of action potentials before and after CNO was calculated for both hM4Di and hM3Dq expressing cells.

Histology

At the conclusion of instrumental training and testing, mice were euthanized and brain tissue was processed to assess viral expression location and spread and GRIN lens placement. Mice from miniscope experiments were anesthetized with isoflurane and transcardially perfused with ice-cold PBS followed by cold 4% paraformaldehyde. The brains were removed, post-fixed in 4% paraformaldehyde, then cryoprotected in 30% sucrose in PBS. 30-μm coronal slices were taken on a cryostat and collected in PBS. Sections were mounted on slides using ProLong Gold antifade reagent with DAPI (Invitrogen). Mice from chemogenetic experiments were anesthetized with isoflurane, brains were removed and rapidly frozen. 20-μm coronal slices were taken on a cryostat and dry mounted onto subbed slides. Slides were submerged in 4% paraformaldehyde, and coverlipped with ProLong Gold antifade reagent with DAPI (Invitrogen). All images were acquired using a Keyence (BZ-X710) microscope with 4X, 10X, and 20X objectives (CFI Plan Apo), CCD camera, and BZ-X Analyze software, to confirm viral expressions and GRIN lens placement.

Statistical Analysis

Datasets were analyzed by 2-tailed t-tests, one-tailed Bayes Factor, or 1-, 2-, or 3-way repeated-measures analysis of variance (ANOVA), or analysis of co-variance (ANCOVA), as appropriate (GraphPad Prism, GraphPad, San Diego, CA; MATLAB, SPSS, IBM, Chicago, IL). ANCOVA controlling for subject was used for all datasets in which cell rather than subject was the unit N. All data sets were checked for normality. Bonferroni post hoc tests corrected for multiple comparisons were performed to clarify statistical interactions. Chemogenetic validation data used uncorrected planned comparisons. Given single-session decoding results, corrected planned comparisons were used for cross-session decoding. Greenhouse-Geisser correction was applied to mitigate the influence of unequal variance between conditions. Alpha levels were set at $P < 0.05$.

Sex as a biological variable

Male and female mice were used in approximately equal numbers, where possible based on breeding and mouse availability, but the *N* per sex was underpowered to examine sex differences. Sex was therefore not included as a factor in statistical analyses, though individual data points are visually disaggregated by sex.

Rigor and reproducibility

See Supplemental Table 4 for key reagents. Group sizes were estimated based on prior work with this behavioral task[101] and to ensure counterbalancing of virus, pellet type, and devaluation test order. Investigators were not blinded to viral group because they were required to administer virus. All behaviors were scored using automated software (Med Associates). Each experiment included at least 1 replication cohort and cohorts were balanced by Viral group or hemisphere (for imaging) prior to the start of the experiment. Investigators were blinded to group when performing histological validation and determining exclusions based on viral spread or mistargeted implant. Calcium analyses included cross-validation and comparison to shuffled data.

Data and code availability

All data that support the findings of this study are available from the corresponding author upon request. Custom-written MATLAB code will be accessible via Dryad repository and available from the corresponding author upon request.

REFERENCES

1 Wassum, K. M. Amygdala-cortical collaboration in reward learning and decision making. *Elife* **11** (2022). https://doi.org/10.7554/eLife.80926

2 Balleine, B. W. The Meaning of Behavior: Discriminating Reflex and Volition in the Brain. *Neuron* **104**, 47-62 (2019). https://doi.org/10.1016/j.neuron.2019.09.024

3 Balleine, B. W. & Dickinson, A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* **37**, 407-419 (1998).

4 O'Doherty, J. P., Cockburn, J. & Pauli, W. M. Learning, Reward, and Decision Making. *Annu Rev Psychol* **68**, 73-100 (2017). https://doi.org/10.1146/annurev-psych-010416-044216

5 Doll, B. B., Simon, D. A. & Daw, N. D. The ubiquity of model-based reinforcement learning. *Curr Opin Neurobiol* **22**, 1075-1081 (2012). https://doi.org/10.1016/j.conb.2012.08.003

6 Dickinson, A. & Balleine, B. W. in *Spatial representation* (eds N. Eilan, R. McCarthy, & M. W. Brewer) 277-293 (Basil Blackwell Ltd, 1993).

7 Dickinson, A. & Balleine, B. W. (1994).

8 Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience* **8**, 1704-1711 (2005).

9 Malvaez, M. & Wassum, K. Regulation of habit formation in the dorsal striatum. *Current Opinion in Behavioral Sciences* **20**, 67-74 (2018).

10 Graybiel, A. M. Habits, rituals, and the evaluative brain. *Annu Rev Neurosci* **31**, 359-387 (2008). https://doi.org/10.1146/annurev.neuro.29.051605.112851

11 Adams, C. D. Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology* **34**, 77-98 (1982).

12 Smith, K. S. & Graybiel, A. M. Habit formation. *Dialogues Clin Neurosci* **18**, 33-43 (2016).

13 Dickinson, A. Actions and Habits: the development of behavioural autonomy. *Philosphical Transactions of the Royal Society of London* **B308**, 67-78 (1985).

14 Dolan, R. J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312-325 (2013). https://doi.org/10.1016/j.neuron.2013.09.007

15 Redish, A. D., Jensen, S. & Johnson, A. A unified framework for addiction: vulnerabilities in the decision process. *Behav Brain Sci* **31**, 415-437; discussion 437-487 (2008). https://doi.org/10.1017/S0140525X0800472X

16 Vandaele, Y. & Ahmed, S. H. Habit, choice, and addiction. *Neuropsychopharmacology* **46**, 689-698 (2021). https://doi.org/10.1038/s41386-020-00899-y

17 Voon, V. *et al.* Disorders of compulsivity: a common bias towards learning habits. *Mol Psychiatry* **20**, 345-352 (2015). https://doi.org/10.1038/mp.2014.44

18 Gillan, C. M., Robbins, T. W., Sahakian, B. J., van den Heuvel, O. A. & van Wingen, G. The role of habit in compulsivity. *Eur Neuropsychopharmacol* **26**, 828-840 (2016). https://doi.org/10.1016/j.euroneuro.2015.12.033

19 Corbit, L. H. & Janak, P. H. Habitual Alcohol Seeking: Neural Bases and Possible Relations to Alcohol Use Disorders. *Alcohol Clin Exp Res* (2016). https://doi.org/10.1111/acer.13094

20 Ostlund, S. B. & Balleine, B. W. On habits and addiction: An associative analysis of compulsive drug seeking. *Drug Discov Today Dis Models* **5**, 235-245 (2008). https://doi.org/10.1016/j.ddmod.2009.07.004

21 Zapata, A., Minney, V. L. & Shippenberg, T. S. Shift from goal-directed to habitual cocaine seeking after prolonged experience in rats. *J Neurosci* **30**, 15457-15463 (2010). https://doi.org/10.1523/JNEUROSCI.4072-10.2010

22 Hogarth, L. & Chase, H. W. Parallel goal-directed and habitual control of human drug-seeking: Implications for dependence vulnerability. *J Exp Psychol Anim Behav Process* (2011).

23 Belin, D., Belin-Rauscent, A., Murray, J. E. & Everitt, B. J. Addiction: failure of control over maladaptive incentive habits. *Curr Opin Neurobiol* (2013). https://doi.org/10.1016/j.conb.2013.01.025

24 Hogarth, L., Balleine, B. W., Corbit, L. H. & Killcross, S. Associative learning mechanisms underpinning the transition from recreational drug use to addiction. *Ann N Y Acad Sci* **1282**, 12-24 (2013). https://doi.org/10.1111/j.1749-6632.2012.06768.x

25 Leblanc, K. H., Maidment, N. T. & Ostlund, S. B. Repeated cocaine exposure facilitates the expression of incentive motivation and induces habitual control in rats. *PLoS One* **8**, e61355 (2013). https://doi.org/10.1371/journal.pone.0061355

26 Furlong, T. M., Jayaweera, H. K., Balleine, B. W. & Corbit, L. H. Binge-like consumption of a palatable food accelerates habitual control of behavior and is dependent on activation of the dorsolateral striatum. *J Neurosci* **34**, 5012-5022 (2014). https://doi.org/10.1523/JNEUROSCI.3707-13.2014

27 Renteria, R., Baltz, E. T. & Gremel, C. M. Chronic alcohol exposure disrupts top-down control over basal ganglia action selection to produce habits. *Nat Commun* **9**, 211 (2018). https://doi.org/10.1038/s41467-017-02615-9

28 Hogarth, L., Attwood, A. S., Bate, H. A. & Munafò, M. R. Acute alcohol impairs human goal-directed action. *Biol Psychol* **90**, 154-160 (2012). https://doi.org/10.1016/j.biopsycho.2012.02.016

29 Ray, L. A. *et al.* Capturing habitualness of drinking and smoking behavior in humans. *Drug Alcohol Depend* **207**, 107738 (2020). https://doi.org/10.1016/j.drugalcdep.2019.107738

30 Groman, S. M., Massi, B., Mathias, S. R., Lee, D. & Taylor, J. R. Model-Free and Model-Based Influences in Addiction-Related Behaviors. *Biol Psychiatry* **85**, 936-945 (2019). https://doi.org/10.1016/j.biopsych.2018.12.017

31 Gillan, C. M. *et al.* Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *Am J Psychiatry* **168**, 718-726 (2011). https://doi.org/10.1176/appi.ajp.2011.10071062

32 Gillan, C. M. *et al.* Enhanced avoidance habits in obsessive-compulsive disorder. *Biol Psychiatry* **75**, 631-638 (2014). https://doi.org/10.1016/j.biopsych.2013.02.002

33 Vaghi, M. M. *et al.* Action-Outcome Knowledge Dissociates From Behavior in Obsessive-Compulsive Disorder Following Contingency Degradation. *Biol Psychiatry Cogn Neurosci Neuroimaging* **4**, 200-209 (2019). https://doi.org/10.1016/j.bpsc.2018.09.014

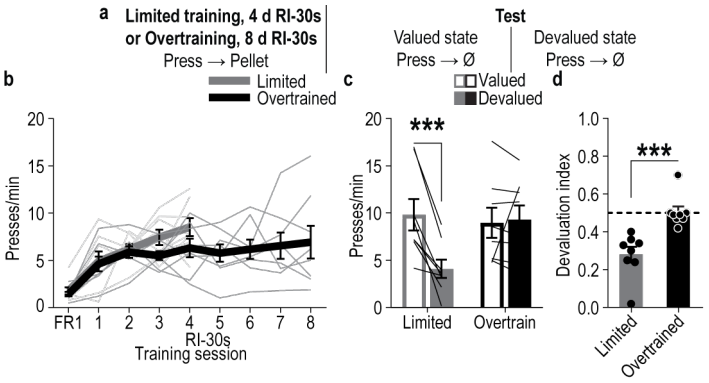
34 Horstmann, A. *et al.* Slave to habit? Obesity is associated with decreased behavioural sensitivity to reward devaluation. *Appetite* **87**, 175-183 (2015). https://doi.org/10.1016/j.appet.2014.12.212

35 Morris, R. W., Cyrzon, C., Green, M. J., Le Pelley, M. E. & Balleine, B. W. Impairments in action-outcome learning in schizophrenia. *Transl Psychiatry* **8**, 54 (2018). https://doi.org/10.1038/s41398-018-0103-0
36 Griffiths, K. R., Morris, R. W. & Balleine, B. W. Translational studies of goal-directed action as a framework for classifying deficits across psychiatric disorders. *Front Syst Neurosci* **8**, 101 (2014). https://doi.org/10.3389/fnsys.2014.00101
37 Morris, R. W., Quail, S., Griffiths, K. R., Green, M. J. & Balleine, B. W. Corticostriatal control of goal-directed action is impaired in schizophrenia. *Biol Psychiatry* **77**, 187-195 (2015). https://doi.org/10.1016/j.biopsych.2014.06.005
38 Byrne, K. A., Six, S. G. & Willis, H. C. Examining the effect of depressive symptoms on habit formation and habit-breaking. *J Behav Ther Exp Psychiatry* **73**, 101676 (2021). https://doi.org/10.1016/j.jbtep.2021.101676
39 Alvares, G. A., Balleine, B. W. & Guastella, A. J. Impairments in goal-directed actions predict treatment response to cognitive-behavioral therapy in social anxiety disorder. *PLoS One* **9**, e94778 (2014). https://doi.org/10.1371/journal.pone.0094778
40 Alvares, G. A., Balleine, B. W., Whittle, L. & Guastella, A. J. Reduced goal-directed action control in autism spectrum disorder. *Autism Res* (2016). https://doi.org/10.1002/aur.1613
41 Malvaez, M. Neural substrates of habit. *J Neurosci Res* (2019). https://doi.org/10.1002/jnr.24552
42 Balleine, B. W. & O'Doherty, J. P. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* **35**, 48-69 (2010). https://doi.org/10.1038/npp.2009.131
43 Balleine, B. W., Delgado, M. R. & Hikosaka, O. The role of the dorsal striatum in reward and decision-making. *J Neurosci* **27**, 8161-8165 (2007). https://doi.org/10.1523/JNEUROSCI.1554-07.2007
44 Yin, H. H., Knowlton, B. J. & Balleine, B. W. Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *Eur J Neurosci* **22**, 505-512 (2005).
45 Corbit, L. H. & Janak, P. H. Posterior dorsomedial striatum is critical for both selective instrumental and Pavlovian reward learning. *Eur J Neurosci* **31**, 1312-1321 (2010). https://doi.org/10.1111/j.1460-9568.2010.07153.x
46 Lex, B. & Hauber, W. The role of dopamine in the prelimbic cortex and the dorsomedial striatum in instrumental conditioning. *Cereb Cortex* **20**, 873-883 (2010). https://doi.org/10.1093/cercor/bhp151
47 Liljeholm, M., Tricomi, E., O'Doherty, J. P. & Balleine, B. W. Neural correlates of instrumental contingency learning: differential effects of action-reward conjunction and disjunction. *J Neurosci* **31**, 2474-2480 (2011). https://doi.org/10.1523/JNEUROSCI.3354-10.2011
48 McNamee, D., Liljeholm, M., Zika, O. & O'Doherty, J. P. Characterizing the associative content of brain structures involved in habitual and goal-directed actions in humans: a multivariate FMRI study. *J Neurosci* **35**, 3764-3771 (2015). https://doi.org/10.1523/JNEUROSCI.4677-14.2015
49 O'Hare, J., Calakos, N. & Yin, H. H. Recent Insights into Corticostriatal Circuit Mechanisms underlying Habits: Invited review for Current Opinions in Behavioral Sciences. *Curr Opin Behav Sci* **20**, 40-46 (2018). https://doi.org/10.1016/j.cobeha.2017.10.001
50 Gremel, C. M. & Costa, R. M. Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nat Commun* **4**, 2264 (2013). https://doi.org/10.1038/ncomms3264
51 Corbit, L. H., Nie, H. & Janak, P. H. Habitual alcohol seeking: time course and the contribution of subregions of the dorsal striatum. *Biol Psychiatry* **72**, 389-395 (2012). https://doi.org/10.1016/j.biopsych.2012.02.024
52 Yin, H. H., Ostlund, S. B., Knowlton, B. J. & Balleine, B. W. The role of the dorsomedial striatum in instrumental conditioning. *Eur J Neurosci* **22**, 513-523 (2005).
53 Peak, J., Hart, G. & Balleine, B. W. From learning to action: the integration of dorsal striatal input and output pathways in instrumental conditioning. *Eur J Neurosci* **49**, 658-671 (2019). https://doi.org/10.1111/ejn.13964
54 Surmeier, D. J., Ding, J., Day, M., Wang, Z. & Shen, W. D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. *Trends Neurosci* **30**, 228-235 (2007). https://doi.org/10.1016/j.tins.2007.03.008
55 Durieux, P. F. *et al.* D2R striatopallidal neurons inhibit both locomotor and drug reward processes. *Nat Neurosci* **12**, 393-395 (2009). https://doi.org/10.1038/nn.2286
56 Albin, R. L., Young, A. B. & Penney, J. B. The functional anatomy of basal ganglia disorders. *Trends Neurosci* **12**, 366-375 (1989).
57 Fang, L. Z. & Creed, M. C. Updating the striatal-pallidal wiring diagram. *Nat Neurosci* **27**, 15-27 (2024). https://doi.org/10.1038/s41593-023-01518-x
58 Graybiel, A. M. Surprises From the Basal Ganglia: Stop and Go Have New Meaning. *Mov Disord* (2025). https://doi.org/10.1002/mds.70008
59 Varin, C. & de Kerchove d'Exaerde, A. Neuronal encoding of behaviors and instrumental learning in the dorsal striatum. *Trends Neurosci* **48**, 77-91 (2025). https://doi.org/10.1016/j.tins.2024.11.003
60 Aharoni, D. & Hoogland, T. M. Circuit Investigations With Open-Source Miniaturized Microscopes: Past, Present and Future. *Front Cell Neurosci* **13**, 141 (2019). https://doi.org/10.3389/fncel.2019.00141
61 Dana, H. *et al.* High-performance calcium sensors for imaging activity in neuronal populations and microcompartments. *Nat Methods* **16**, 649-657 (2019). https://doi.org/10.1038/s41592-019-0435-6
62 Adams, C. D. & Dickinson, A. Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology* **33**, 109-121 (1981).
63 Cazares, C., Schreiner, D. C., Valencia, M. L. & Gremel, C. M. Orbitofrontal cortex populations are differentially recruited to support actions. *Curr Biol* **32**, 4675-4687.e4675 (2022). https://doi.org/10.1016/j.cub.2022.09.022
64 Schreiner, D. C., Cazares, C., Renteria, R. & Gremel, C. M. Information normally considered task-irrelevant drives decision-making and affects premotor circuit recruitment. *Nat Commun* **13**, 2134 (2022). https://doi.org/10.1038/s41467-022-29807-2
65 Gu, X. & Johansen, J. P. Prefrontal encoding of an internal model for emotional inference. *Nature* **643**, 1044-1056 (2025). https://doi.org/10.1038/s41586-025-09001-2
66 Jin, X. & Costa, R. M. Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature* **466**, 457-462 (2010). https://doi.org/nature09263 [pii]10.1038/nature09263
67 Smith, K. S. & Graybiel, A. M. A dual operator view of habitual behavior reflecting cortical and striatal dynamics. N*euron* 7**9,** 361-374 (2013). https://doi.org/10.1016/j.neuron.2013.05.038

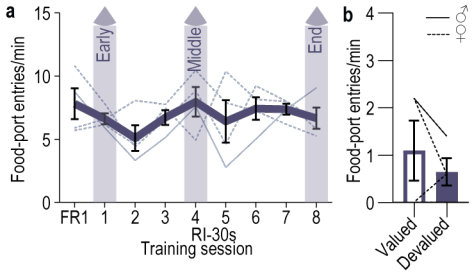
68 Lee, J. & Sabatini, B. L. Striatal indirect pathway mediates exploration via collicular competition. N*ature* 5**99,** 645-649 (2021). https://doi.org/10.1038/s41586-021-04055-4
69 Kravitz, A. V., Tye, L. D. & Kreitzer, A. C. Distinct roles for direct and indirect pathway striatal neurons in reinforcement. N*at Neurosci* 1**5,** 816-818 (2012). https://doi.org/10.1038/nn.3100
70 Yttri, E. A. & Dudman, J. T. Opponent and bidirectional control of movement velocity in the basal ganglia. N*ature* 5**33,** 402-406 (2016). https://doi.org/10.1038/nature17639
71 Tecuapetla, F., Jin, X., Lima, S. Q. & Costa, R. M. Complementary Contributions of Striatal Projection Pathways to Action Initiation and Execution. C*ell* (2016).
72 Kravitz, A. V. *et al.* Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. N*ature* 4**66,** 622-626 (2010). https://doi.org/10.1038/nature09159
73 Freeze, B. S., Kravitz, A. V., Hammack, N., Berke, J. D. & Kreitzer, A. C. Control of basal ganglia output by direct and indirect pathway projection neurons. J *Neurosci* 3**3,** 18531-18539 (2013). https://doi.org/10.1523/JNEUROSCI.1278-13.2013
74 Pomrenze, M. B. *et al.* A Transgenic Rat for Investigating the Anatomy and Function of Corticotrophin Releasing Factor Circuits. F*ront Neurosci* 9**,** 487 (2015). https://doi.org/10.3389/fnins.2015.00487
75 Tipps, M., Marron Fernandez de Velasco, E., Schaeffer, A. & Wickman, K. Inhibition of Pyramidal Neurons in the Basal Amygdala Promotes Fear Learning. e*Neuro* 5 (2018). https://doi.org/10.1523/ENEURO.0272-18.2018
76 Tuscher, J. J., Taxier, L. R., Fortress, A. M. & Frick, K. M. Chemogenetic inactivation of the dorsal hippocampus and medial prefrontal cortex, individually and concurrently, impairs object recognition and spatial memory consolidation in female mice. N*eurobiol Learn Mem* 1**56,** 103-116 (2018). https://doi.org/10.1016/j.nlm.2018.11.002
77 Alexander, G. M. *et al.* Remote control of neuronal activity in transgenic mice expressing evolved G protein-coupled receptors. N*euron* 6**3,** 27-39 (2009). https://doi.org/10.1016/j.neuron.2009.06.014
78 Vazey, E. M. & Aston-Jones, G. Designer receptor manipulations reveal a role of the locus coeruleus noradrenergic system in isoflurane general anesthesia. P*roc Natl Acad Sci U S A* 1**11,** 3859-3864 (2014). https://doi.org/10.1073/pnas.1310025111
79 Qiu, M. H., Chen, M. C., Fuller, P. M. & Lu, J. Stimulation of the Pontine Parabrachial Nucleus Promotes Wakefulness via Extra-thalamic Forebrain Circuit Nodes. C*urr Biol* 2**6,** 2301-2312 (2016). https://doi.org/10.1016/j.cub.2016.07.054
80 Zhu, H. *et al.* Cre-dependent DREADD (Designer Receptors Exclusively Activated by Designer Drugs) mice. G*enesis* 5**4,** 439-446 (2016). https://doi.org/10.1002/dvg.22949
81 Shan, Q., Ge, M., Christie, M. J. & Balleine, B. W. The acquisition of goal-directed actions generates opposing plasticity in direct and indirect pathways in dorsomedial striatum. J *Neurosci* 3**4,** 9196-9201 (2014). https://doi.org/10.1523/JNEUROSCI.0313-14.2014
82 Yu, X., Chen, S. & Shan, Q. Depression in the Direct Pathway of the Dorsomedial Striatum Permits the Formation of Habitual Action. C*ereb Cortex* 3**1,** 3551-3564 (2021). https://doi.org/10.1093/cercor/bhab031
83 Peak, J., Chieng, B., Hart, G. & Balleine, B. W. Striatal direct and indirect pathway neurons differentially control the encoding and updating of goal-directed learning. E*life* 9 (2020). https://doi.org/10.7554/eLife.58544
84 Ferguson, S. M., Phillips, P. E., Roth, B. L., Wess, J. & Neumaier, J. F. Direct-pathway striatal neurons regulate the retention of decision-making strategies. J *Neurosci* 3**3,** 11668-11676 (2013). https://doi.org/10.1523/JNEUROSCI.4783-12.2013
85 Gerfen, C. R. & Surmeier, D. J. Modulation of striatal projection systems by dopamine. A*nnu Rev Neurosci* 3**4,** 441-466 (2011). https://doi.org/10.1146/annurev-neuro-061010-113641
86 Nonomura, S. *et al.* Monitoring and Updating of Action Selection for Goal-Directed Behavior through the Striatal Direct and Indirect Pathways. N*euron* 9**9,** 1302-1314.e1305 (2018). https://doi.org/10.1016/j.neuron.2018.08.002
87 Cui, G. *et al.* Concurrent activation of striatal direct and indirect pathways during action initiation. N*ature* 4**94,** 238-242 (2013). https://doi.org/10.1038/nature11846
88 Cox, J. & Witten, I. B. Striatal circuits for reward learning and decision-making. N*at Rev Neurosci* 2**0,** 482-494 (2019). https://doi.org/10.1038/s41583-019-0189-2
89 Delevich, K. *et al.* Activation, but not inhibition, of the indirect pathway disrupts choice rejection in a freely moving, multiple-choice foraging task. C*ell Rep* 4**0,** 111129 (2022). https://doi.org/10.1016/j.celrep.2022.111129
90 Jin, X., Tecuapetla, F. & Costa, R. M. Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences. N*at Neurosci* 1**7,** 423-430 (2014). https://doi.org/10.1038/nn.3632
91 Ramírez-Armenta, K. I. *et al.* Optogenetic inhibition of indirect pathway neurons in the dorsomedial striatum reduces excessive grooming in Sapap3-knockout mice. N*europsychopharmacology* 4**7,** 477-487 (2022). https://doi.org/10.1038/s41386-021-01161-9
92 Bakhurin, K. I. *et al.* Opponent regulation of action performance and timing by striatonigral and striatopallidal pathways. E*life* 9 (2020). https://doi.org/10.7554/eLife.54831
93 Carvalho Poyraz, F. *et al.* Decreasing Striatopallidal Pathway Function Enhances Motivation by Energizing the Initiation of Goal-Directed Action. J *Neurosci* 3**6,** 5988-6001 (2016). https://doi.org/10.1523/JNEUROSCI.0444-16.2016
94 Li, Y. *et al.* Optogenetic Activation of Adenosine A2A Receptor Signaling in the Dorsomedial Striatopallidal Neurons Suppresses Goal-Directed Behavior. N*europsychopharmacology* 4**1,** 1003-1013 (2016). https://doi.org/10.1038/npp.2015.227
95 Tecuapetla, F., Koós, T., Tepper, J. M., Kabbani, N. & Yeckel, M. F. Differential dopaminergic modulation of neostriatal synaptic connections of striatopallidal axon collaterals. J *Neurosci* 2**9,** 8977-8990 (2009). https://doi.org/10.1523/JNEUROSCI.6145-08.2009
96 Furlong, T. M., Supit, A. S., Corbit, L. H., Killcross, S. & Balleine, B. W. Pulling habits out of rats: adenosine 2A receptor antagonism in dorsomedial striatum rescues meth-amphetamine-induced deficits in goal-directed action. A*ddict Biol* (2015). https://doi.org/10.1111/adb.12316
97 Matamales, M. *et al.* Local D2- to D1-neuron transmodulation updates goal-directed learning in the striatum. S*cience* 3**67,** 549-555 (2020). https://doi.org/10.1126/science.aaz5751
98 Gagnon, D. *et al.* Striatal neurons expressing D1 and D2 receptors are morphologically distinct and differently affected by dopamine denervation in mice. S*ci Rep* 7**,** 41432 (2017). https://doi.org/10.1038/srep41432

99 Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174,** 1015-1030.e1016 (2018). https://doi.org/10.1016/j.cell.2018.07.028
100 Gokce, O. *et al.* Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell RNA-Seq. *Cell Rep* **16,** 1126-1137 (2016). https://doi.org/10.1016/j.celrep.2016.06.059
101 Malvaez, M. *et al.* Habits Are Negatively Regulated by Histone Deacetylase 3 in the Dorsal Striatum. *Biol Psychiatry* (2018). https://doi.org/10.1016/j.biopsych.2018.01.025
102 Shiflett, M. W. & Balleine, B. W. Molecular substrates of action control in cortico-striatal circuits. *Prog Neurobiol* **95,** 1-13 (2011). https://doi.org/S0301-0082(11)00085-2 [pii]10.1016/j.pneurobio.2011.05.007
103 Gremel, C. M. *et al.* Endocannabinoid Modulation of Orbitostriatal Circuits Gates Habit Formation. *Neuron* **90,** 1312-1324 (2016). https://doi.org/10.1016/j.neuron.2016.04.043
104 Fisher, S. D., Ferguson, L. A., Bertran-Gonzalez, J. & Balleine, B. W. Amygdala-Cortical Control of Striatal Plasticity Drives the Acquisition of Goal-Directed Action. *Curr Biol* (2020). https://doi.org/10.1016/j.cub.2020.08.090
105 Giovanniello, J. *et al.* A dual-pathway architecture for stress to disrupt agency and promote habit. *Nature* (In Press).
106 Dias-Ferreira, E. *et al.* Chronic stress causes frontostriatal reorganization and affects decision-making. *Science* **325,** 621-625 (2009). https://doi.org/10.1126/science.1171203
107 Giovanniello, J. R. *et al.* A dual-pathway architecture for stress to disrupt agency and promote habit. *Nature* **640,** 722-731 (2025). https://doi.org/10.1038/s41586-024-08580-w
108 Seiler, J. L. *et al.* Dopamine signaling in the dorsomedial striatum promotes compulsive behavior. *Curr Biol* **32,** 1175-1188.e1175 (2022). https://doi.org/10.1016/j.cub.2022.01.055
109 Hopf, F. W., Mailliard, W. S., Gonzalez, G. F., Diamond, I. & Bonci, A. Atypical protein kinase C is a novel mediator of dopamine-enhanced firing in nucleus accumbens neurons. *J Neurosci* **25,** 985-989 (2005).
110 Hart, G., Burton, T. J., Nolan, C. R. & Balleine, B. W. Striatal dopamine release tracks the relationship between actions and their consequences. *Cell Rep* **43,** 113828 (2024). https://doi.org/10.1016/j.celrep.2024.113828
111 Bradfield, L. A., Bertran-Gonzalez, J., Chieng, B. & Balleine, B. W. The thalamostriatal pathway and cholinergic control of goal-directed action: interlacing new with existing learning in the striatum. *Neuron* **79,** 153-166 (2013). https://doi.org/10.1016/j.neuron.2013.04.039
112 Kang, S. *et al.* Activation of Astrocytes in the Dorsomedial Striatum Facilitates Transition From Habitual to Goal-Directed Reward-Seeking Behavior. *Biol Psychiatry* **88,** 797-808 (2020). https://doi.org/10.1016/j.biopsych.2020.04.023
113 Dobbs, L. K. *et al.* Dopamine Regulation of Lateral Inhibition between Striatal Neurons Gates the Stimulant Actions of Cocaine. *Neuron* **90,** 1100-1113 (2016). https://doi.org/10.1016/j.neuron.2016.04.031
114 Taverna, S., Ilijic, E. & Surmeier, D. J. Recurrent collateral connections of striatal medium spiny neurons are disrupted in models of Parkinson's disease. *J Neurosci* **28,** 5504-5512 (2008). https://doi.org/10.1523/JNEUROSCI.5493-07.2008
115 Tai, L. H., Lee, A. M., Benavidez, N., Bonci, A. & Wilbrecht, L. Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nat Neurosci* **15,** 1281-1289 (2012). https://doi.org/10.1038/nn.3188
116 Stalnaker, T. A., Calhoon, G. G., Ogawa, M., Roesch, M. R. & Schoenbaum, G. Neural correlates of stimulus-response and response-outcome associations in dorsolateral versus dorsomedial striatum. *Front Integr Neurosci* **4**, 12 (2010). https://doi.org/10.3389/fnint.2010.00012
117 Yin, H. H., Park, B. S., Adermark, L. & Lovinger, D. M. Ethanol reverses the direction of long-term synaptic plasticity in the dorsomedial striatum. *Eur J Neurosci* **25,** 3226-3232 (2007). https://doi.org/10.1111/j.1460-9568.2007.05606.x
118 Hayrapetyan, V. *et al.* Region-specific impairments in striatal synaptic transmission and impaired instrumental learning in a mouse model of Angelman syndrome. *Eur J Neurosci* **39,** 1018-1025 (2014). https://doi.org/10.1111/ejn.12442
119 Bertran-Gonzalez, J., Dinale, C. & Matamales, M. Restoring the youthful state of striatal plasticity in aged mice re-enables cognitive control of action. *Curr Biol* **33,** 1997-2007.e1995 (2023). https://doi.org/10.1016/j.cub.2023.04.020
120 Lemberger, T. *et al.* Expression of Cre recombinase in dopaminoceptive neurons. *BMC Neurosci* **8**, 4 (2007). https://doi.org/10.1186/1471-2202-8-4
121 Gong, S. *et al.* Targeting Cre recombinase to specific neuron populations with bacterial artificial chromosome constructs. *J Neurosci* **27,** 9817-9823 (2007). https://doi.org/10.1523/JNEUROSCI.2707-07.2007
122 Blair, G. J. *et al.* Hippocampal place cell remapping occurs with memory storage of aversive experiences. *Elife* **12** (2023). https://doi.org/10.7554/eLife.80661
123 Giovannucci, A. *et al.* CaImAn an open source tool for scalable calcium imaging data analysis. *Elife* **8** (2019). https://doi.org/10.7554/eLife.38173
124 Pnevmatikakis, E. A. Analysis pipelines for calcium imaging data. *Curr Opin Neurobiol* **55,** 15-21 (2019). https://doi.org/10.1016/j.conb.2018.11.004
125 Sheintuch, L. *et al.* Tracking the Same Neurons across Multiple Days in Ca. *Cell Rep* **21,** 1102-1115 (2017). https://doi.org/10.1016/j.celrep.2017.10.013
126 Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **16** (2002).
127 Ding, J. *et al.* An expanded GCaMP reporter toolkit for functional imaging in Caenorhabditis elegans. *G3 (Bethesda)* **13** (2023). https://doi.org/10.1093/g3journal/jkad183

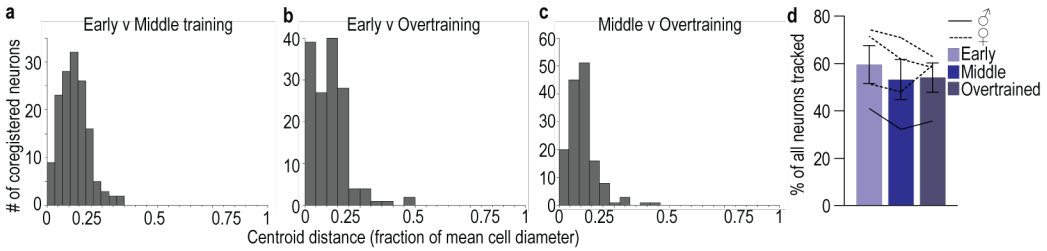
EXTENDED DATA FIGURES



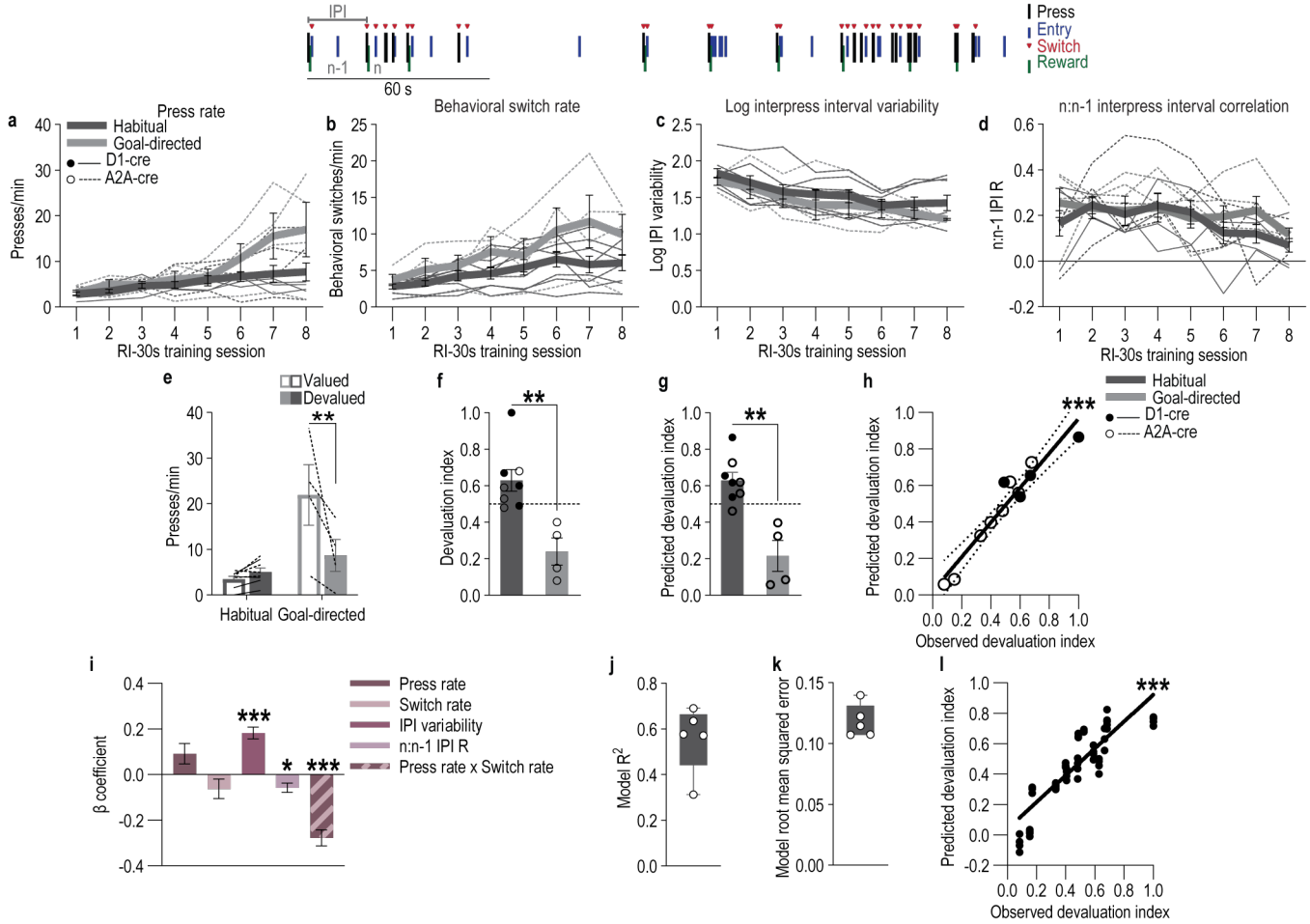
Extended Data Figure 1-1: Behavior is goal-directed following limited random-interval instrumental training and habitual following overtraining. (a) Procedure. Lever presses earned food pellet rewards on a random-interval (RI) 30-s reinforcement schedule. Mice received either limited training (4 RI sessions) or overtraining (8 sessions). Mice were then given a lever-pressing probe test in the Valued state, prefed on untrained food-pellet type to control for general satiety and Devalued state prefed on trained food-pellet type to induce sensory-specific satiety devaluation. Test order was counterbalanced across subjects within each group, with a single intervening retraining session. **(b)** Training press rate. 1-way ANOVA, Limited training: Training: $F_{2.75,\ 19.27} = 12.08$, $P = 0.0001$. Overtraining: Training: $F_{2.24,\ 15.69} = 4.56$, $P = 0.02$. **(c)** Test press rate. 2-way ANOVA, Value x Training duration: $F_{1,\ 14} = 14.69$, $P = 0.002$; Value: $F_{1,\ 14} = 11.69$, $P = 0.004$; Training duration: $F_{1,\ 14} = 1.31$, $P = 0.27$. **(d)** Devaluation index [(Devalued presses)/(Valued presses + Devalued presses)]. 2-tailed Mann-Whitney U test, U = 0, P = 0.002. N = 8/group (all male). Data presented as mean ± s.e.m. ***$P < 0.001$. Mice learn action-outcome relationships during instrumental conditioning on a random-interval schedule of reinforcement and use them for goal-directed decision making after limited training and form habits with overtraining.



Extended Data Figure 1-2: Entries into the food-delivery port during training and test for DMS D1⁺ imaging experiment. (a) Training entry rate. 1-way ANOVA, Training: $F_{2.11, 6.31} = 0.75$, $P = 0.52$ **(b)** Test entry rate. 2-tailed t-test, $t_3 = 0.94$, $P = 0.42$, 95% CI -1.97 - 1.07. D1-cre: $N = 4$ (1 male). Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines.

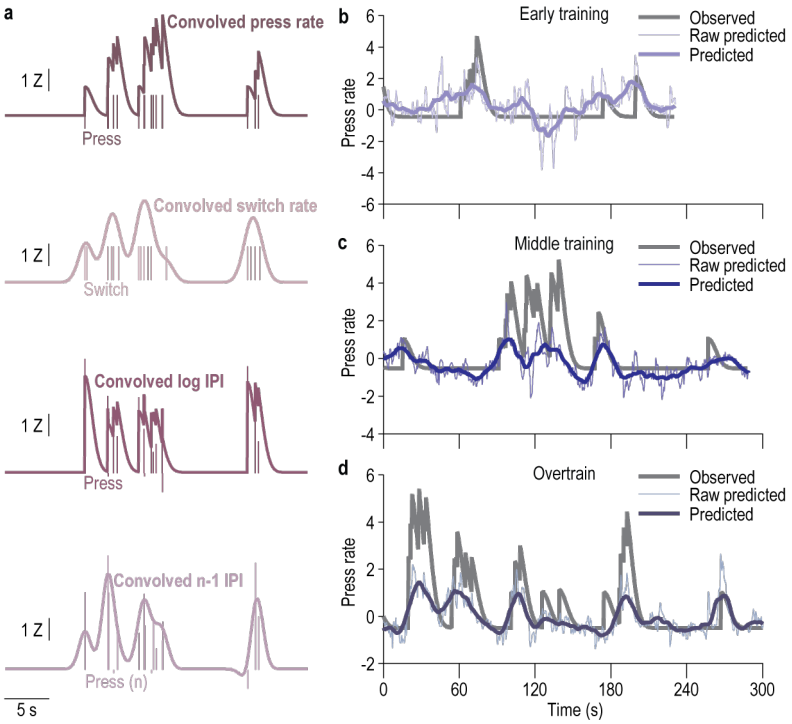


Extended Data Figure 1-3. Coregistration of DMS D1⁺ neurons across training. (a-c) Representative example of coregistration of DMS D1⁺ neurons across training. Distribution of the distance between cell centroids (as a fraction of total cell diameter) of co-registered cell pairs in the 1st (a, early) and 4th (middle), 1st and 8th (b, overtrain), and the 4th and 8th (c) training sessions. **(d)** Percent of all DMS D1⁺ neurons coregistered across training. 1-way ANOVA, $F_{1.19, 3.58} = 1.62$, $P = 0.29$. D1-cre: $N = 4$ (1 male). We were able to coregister on average 56% (s.e.m. = 3.90%) of the D1⁺ neurons across training. Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines.

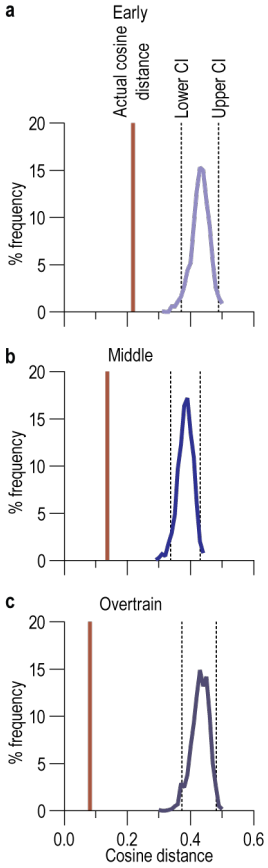


Extended Data Figure 1-4. Linear regression model of goal-directed v. habitual behavioral strategy. To reveal behavioral variables that relate to the extent of goal-directed v. habitual behavioral strategy we used a linear regression model to ask whether and how behavior during the final training session predicts subsequent sensitivity to outcome devaluation. We pooled behavioral data from all the D1-cre and A2A-cre subjects (N = 12) so we could exploit the variability in sensitivity to devaluation for this model. We included measures of behavioral engagement: lever-press rate and switch rate (rate at which subject switch from pressing to checking the food port or vice versa). We also included measures of behavioral organization. On the random-interval reinforcement schedule a variable amount of time must elapse after an earned reward before another press will earn a reward. Thus, we focused our organization measures of the time subjects waited in between presses (interpress interval, IPI). As a measure of overall organization, we included the variability/consistency (standard deviation) of the log-transformed interpress intervals. As a measure of short-term patterning and the extent to which current trial behavior relates to past trial behavior, we included the correlation between the current interpress interval, n, and the interpress interval of the prior press, n-1[63,64]. Schematic shows a representative example of these behavioral events across time. **(a-d)** Model predictor variables across training. Data are divided by subjects that were sensitive to devaluation ('Goal-directed', devaluation index < group mean of 0.483) v. those that were insensitive ('Habitual', devaluation index > 0.483). Lever-press rate (a; 2-way ANOVA, Training: $F_{1.66, 16.55}$ = 11.59, P = 0.001; Devaluation sensitivity group: $F_{1, 10}$ = 2.72, P = 0.13; Training x Group: $F_{1.66, 16.55}$ = 3.36, P = 0.07), behavioral switch rate (b; 2-way ANOVA, Training: $F_{2.89, 28.89}$ = 12.99, P < 0.0001; Devaluation sensitivity group: $F_{1, 10}$ = 2.68, P = 0.13; Training x Group: $F_{2.89, 28.89}$ = 2.17, P = 0.11), variability/consistency (standard deviation) of log interpress intervals (c; 2-way ANOVA, Training: $F_{2.67, 26.70}$ = 12.29, P < 0.0001; Devaluation sensitivity group: $F_{1, 10}$ = 0.86, P = 0.38; Training x Group: $F_{2.67, 26.70}$ = 0.37, P = 0.75), and relationship to past trial interpress interval (n:n-1 interpress interval correlation; d; 2-way ANOVA, Training: $F_{2.86, 28.61}$ = 13.10, P < 0.0001; Devaluation sensitivity group: $F_{1, 10}$ = 1.70, P = 0.22; Training x Group: $F_{2.86, 28.61}$ = 1.13, P = 0.35). Behavioral engagement and transitions tended to increase with training. Interpress-interval variability decreased with training, indicating more consistent overall interpress interval organization. But short-term patterning, the relationship between current trial press wait times and past trial wait, tended to decrease with training, indicating reduced use of past behavior to inform current behavior. These effects did not significantly differ for subjects for which behavior was

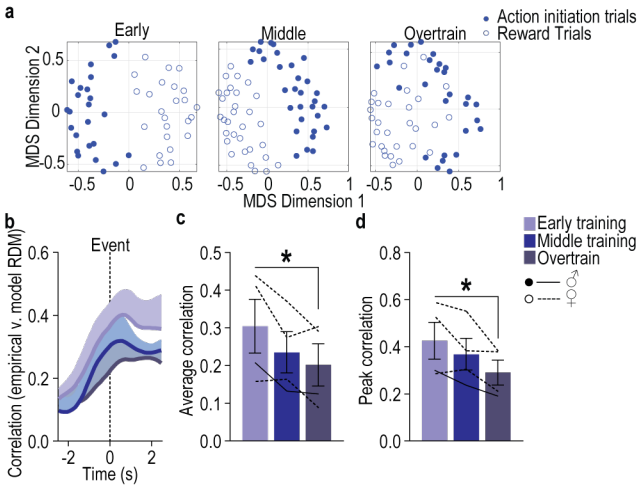
primarily goal directed v. those for which it was primarily habitual. **(e-f)** Devaluation test press rate (e; 2-way ANOVA, Devaluation sensitivity group x Devaluation: $F_{1, 10} = 8.74$, $P = 0.01$; Group: $F_{1, 10} = 12.18$, $P = 0.006$; Devaluation: $F_{1, 10} = 13.88$, $P = 0.004$ and devaluation index (f; 2-tailed t-test with Welch's correction, $t_{6.77} = 4.09$, $P = 0.005$, 95% CI -0.62 - -0.16). **(g-i)** Linear regression model of devaluation index. Using the Matlab stepwiselm function, we identified the model including press rate, switch rate, log interpress-interval variability, and n:n-1 interpress interval correlation that provided the best fit for devaluation index. In preliminary investigations, we also evaluated whether the influence of the interpress interval variables on devaluation sensitivity depended on whether there was a food-port check during the interval and found it did not. The best model (*DevaluationIndex ~ 1 + StdevlogIPI + IPIn:IPIn-1corr + PressRate*SwitchRate*) provided a strong fit of the data (AIC = -20.59, BIC = -16.12, Log-Likelihood = 17.30, and Deviance = -34.59) and significantly predicted devaluation index (β = 0.70, SE = 0.03; $t_8 = 20.44$, $P < 0.0001$, 95% CI 0.62 – 0.77). (g) Devaluation index predicted by the linear regression model (2-tailed t-test with Welch's correction, $t_{4.67} = 4.33$, $P = 0.009$, 95% CI -0.67 – -0.16). (h) Correlation between actual and model-predicted devaluation index (Pearson $r_{12} = 0.96$, $P < 0.0001$, 95% CI 0.86 – 0.99). (i) Individual variable β coefficients. Neither press rate (β = 0.09, SE = 0.04, $t_8 = 2.06$, $P = 0.07$, 95% CI -0.01 – 0.19) nor switch rate (β = -0.06, SE = 0.04; $t_8 = -1.47$, $P = 0.18$, 95% CI -0.16 – 0.03) alone significantly predicted devaluation index, but the interaction between these was a significant negative predictor (β = -0.28, SE = 0.04, $t_8 = -7.86$, $P < 0.0001$, 95% CI -0.36 – 0.20), indicating that as switch rate increases, press rate negatively predicts devaluation index. When press rates are high and animals are also flexibly switching between pressing and checking the food-delivery port this is associated with goal-directed behavioral control (sensitivity to devaluation). When press rates are high, but animals are being inflexible (e.g., continuing to press the lever without checking the food port), this is associated with habit (insensitivity to devaluation). Interpress-interval variability was a significant positive predictor of devaluation index (β = 0.18, SE = 0.03, $t_8 = 7.08$, $P = 0.0001$, 95% CI 0.12 – 0.24), indicating that higher variability in the time between presses (high standard deviation of interpress intervals) predicts habitual behavioral strategy (insensitivity to devaluation). Intuitively, habits might be thought to be less variable. However, this result indicates that the extent of goal-directed v. habitual behavioral strategy is predicted by how strongly the time an animal waits between presses relates to overall organization of waiting. Indeed, the extent to which current trial press wait time relates to past trial behavior was a significant negative predictor of devaluation index (β = -0.06, SEM = 0.02, $t_8 = -2.78$, $P = 0.02$, 95% CI -0.10 – -0.01). A stronger relationship between current interpress intervals and past interpress intervals predicted goal-directed behavioral strategy (sensitivity to devaluation). Thus, goal-directed actions are more and habits are less informed by how the action has been performed recently. Together this model indicates that goal-directed actions are marked by being on task, engaged in pressing and checking for the outcome and re-engagement in pressing that is planned based on, or at least related to, past performance. Habits, on the other hand, are marked by less task engagement or less flexible task engagement and less consideration of past successful behavior when performing the action. **(j-l)** Model performance was evaluated with a 5-fold cross-validation procedure repeated 5 times. Average adjusted R-squared (j) and root mean square error (k) for each cross-validation. Correlation between predicted and actual devaluation index for each model cross validation (l; Pearson R, $r_{70} = 0.87$, $P < 0.0001$, 95% CI 0.79 – 0.92). D1-cre: $N = 12$ (D1-cre $N = 4$, A2A-cre $N = 8$). Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines. $^{**}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.



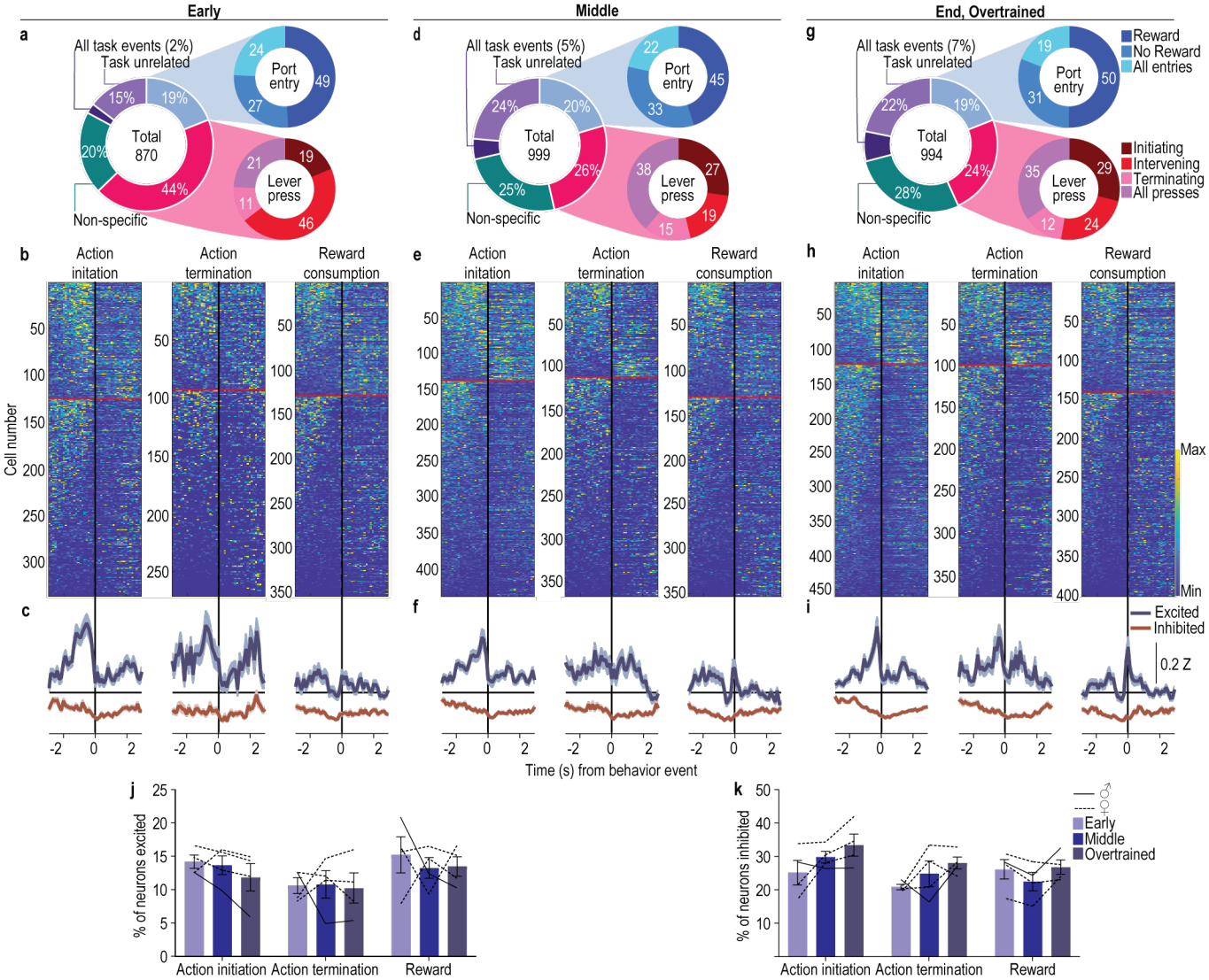
Extended Data Figure 1-5. Representative linear decoding. (a) Example convolved press rate, switch rate, current-trial log interpress interval (IPI), and prior trial interpress interval. Vertical lines: time of each event (press, switch, lever press for which the current log or prior interval is calculated), height represents duration of intervals. **(b-d)** Representative example of decoded press rate from the activity of coregistered DMS D1^{+} neurons across training Actual observed and predicted lever-press rate from the 1st (a, early), 4th (b, middle), and 8th (c, overtrain) training session for a single subject.



Extended Data Figure 1-6. Representative DMS D1⁺ action and reward outcome population vector cosine distance compared to shuffled control. The distribution of cosign distances of a representative subject between action initiation and reward outcome D1⁺ neuronal population vectors for each phase of training (a, early; b, middle; c; overtrain) by bootstrap resampling. The angular distance of the paired group (red line) falls outside of the 95% confidence interval (dashed lines).

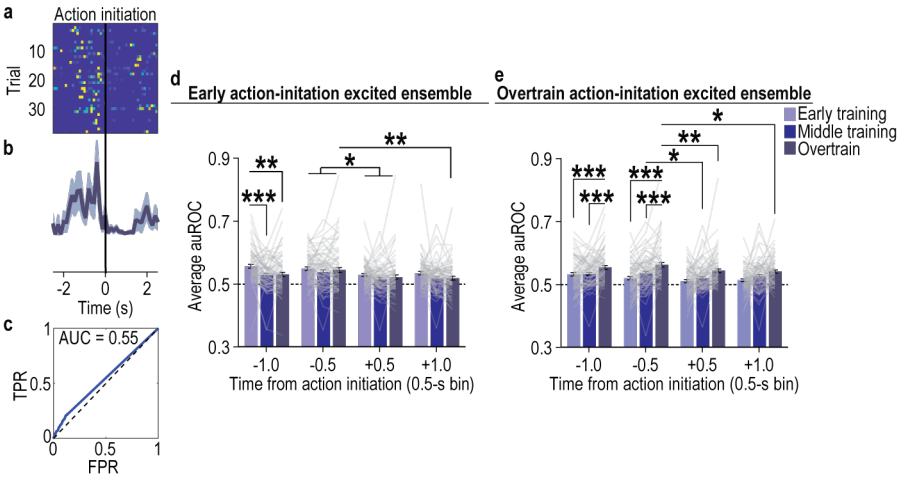


Extended Data Figure 1-7. DMS D1⁺ population representation of action and reward becomes more similar with learning. To assess how the neural activity patterns for action initiation and reward evolve over the course of learning, we used a time-resolved Representational Similarity analysis that allows us to examine the geometry of the neural activity within the entire neural population for each occurrence of action initiation and reward across training session. Whereas the population vector analysis considers mean activity across trials, this approach includes trial x trial consistency and, therefore, provides a complementary analysis to the population vector approach. **(a)** Representative single-subject example multidimensional scaling (MDS) plots of the Representational Dissimilarity Matrix (RDM) of pairwise dissimilarities between the DMS D1⁺ population activity patterns around action initiation and earned reward from the 1st (a, early), 4th (b, middle), and 8th (c, overtrain) training session. **(b-d)** Spearman's rank correlation values between the empirical RDM of the D1⁺ population action initiation and reward activity patterns and those of a theoretical RDM generated to hypothesize complete dissimilarity between action initiation and reward (lower correlation indicates similar neural representations between action initiation and reward) across time centered around each event (b), average (c; 1-way RM ANOVA, $F_{1.54, 4.63} = 8.72$, $P = 0.03$), or peak (d; 1-way RM ANOVA, $F_{1.88, 5.64} = 8.63$, $P = 0.02$) +/- 2.5 s around the events. With learning and continued training the DMS D1⁺ neuronal population representation of action initiation and earned reward becomes more similar, indicating of a representational merging of the action and outcome. D1-cre: $N = 4$ (1 male). Data presented as mean ± s.e.m. Males = closed circles/solid lines, Females = open circles/dashed lines. *$P < 0.05$.

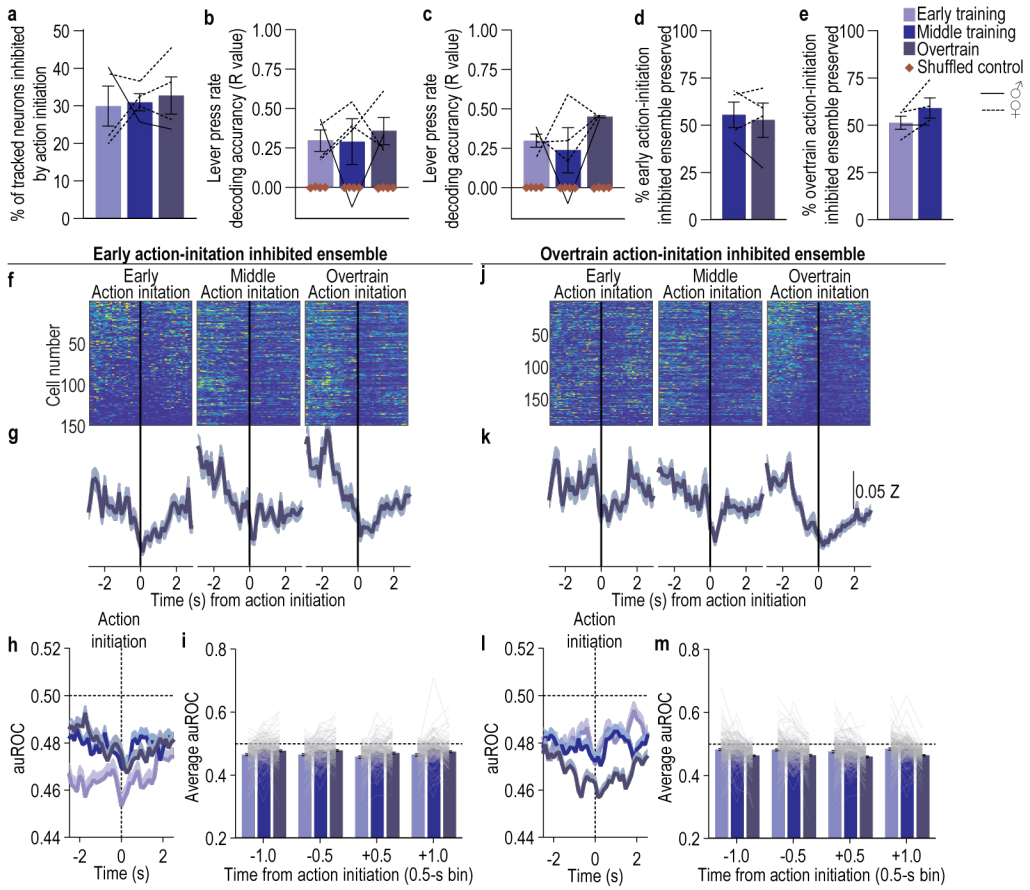


Extended Data Figure 2-1. DMS D1⁺ neurons are modulated by actions and rewards during instrumental learning and overtraining. To ask how DMS D1⁺ neurons encode task events during action-outcome learning and the transition to habit, we computed area under the receiver operating characteristic (auROC) values and classified neurons as significantly modulated by critical behavioral events: lever-press action initiation (first press in the session, after earned reward collection, or after a non-reinforced food-port check), action termination (last press before reward collection or non-reinforced food-port check), intervening lever presses (all other presses), non-reinforced food-port checks, and reward collection. 76 - 85% of the neurons significantly increased or decreased their activity within ±2.9 s around one or more of these events. **(a-c)** Activity of DMS D1⁺ neurons on the 1st (Early) session of RI training. (a) Percent of all recorded neurons (*N* = 870) significantly modulated by lever presses, food-delivery port checks, and reward. (b) Heat map of minimum to maximum deconvolved activity (sorted by total activity) of each D1⁺ neuron significantly modulated around lever-press action initiation (left), action termination (middle), or reward consumption (right). Above red line = excited above cutoff criterion, below = inhibited. (c) Z-scored activity of each population of modulated neurons. **(d-f)** Activity of DMS D1⁺ neurons on the 4th (Middle) training session. (d) Percent of all recorded neurons (*N* = 999) significantly modulated by lever presses, food-delivery port checks, and reward. (e) Heat map of each D1⁺ neuron significantly modulated around lever-press action initiation, action termination, or reward consumption. (f) Z-scored activity of each population of modulated neurons. **(g-i)** Activity of DMS D1⁺ neurons on the 8th (End, Overtrain) training session. (g) Percent of all recorded neurons (*N* = 994) significantly modulated by lever presses, food-delivery port checks, and reward. (h) Heat map of each DMS D1⁺ neuron significantly modulated around lever-press action initiation, action termination, or reward consumption. (i) Z-scored activity of each population of modulated neurons. **(j-k)** Percent of all recorded neurons (Early: *N* = 870, Middle: *N* = 999; End *N* = 994) classified (auROC values >95th percentile of the distribution of shuffled auROCs within 2 s before or after event) as significantly excited (j; 2-way ANOVA,

Training session: $F_{1.07, 3.21} = 0.24$, $P = 0.67$; Event: $F_{1.34, 4.03} = 2.76$, $P = 0.17$; Training x Event: $F_{1.71, 5.14} = 0.48$, $P = 0.62$) or inhibited (k; -way ANOVA, Training: $F_{1.75, 5.26} = 5.34$, $P = 0.06$; Event: $F_{1.47, 4.40} = 4.32$, $P = 0.10$; Training x Event: $F_{1.66, 4.98} = 1.56$, $P = 0.29$) around initiating lever presses, terminating lever presses, or reward consumption. D1-cre: $N = 4$ (1 male). Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines. 76 - 85% of DMS $D1^+$ neurons significantly increased or decreased their activity within ±2.9 s around one or more of task events. 11 - 14% of DMS $D1^+$ neurons were activated around, largely immediately prior to, action initiation. Approximately 10% of neurons were activated around, mostly prior to, action termination. 13 – 15% of neurons responded to the earned reward. Similar proportions of neurons were activated by each task event across training, though there was a slight trend towards more neurons being inhibited around actions with overtraining. Thus, find that DMS $D1^+$ neurons encode actions and outcomes during learning and as habits form with overtraining.

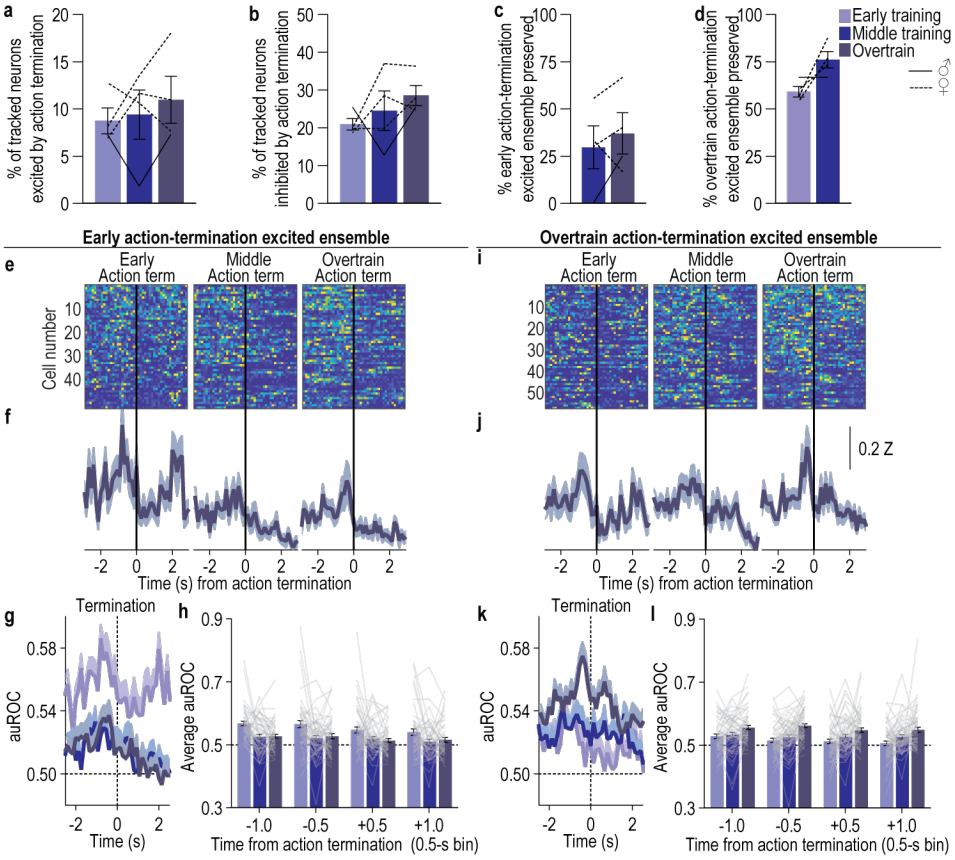


Extended Data Figure 2-2. Example and quantification of the modulation of DMS D1⁺ action-initiation excited neurons. (a-c) Representative single-subject/single-session example of the trial x trial activity of a single action-initiation excited DMS D1⁺ neuron. (a) Heat map of minimum to maximum deconvolved activity sorted by trial of a single D1⁺ action-initiation excited DMS D1⁺ neuron. (b) Z-scored, trial-averaged activity of this neuron. (c) ROC plot with true positive rate (TPR) and false positive rate (FPR) for this neuron compared to chance level descrimination (dashed line), with the area uder the ROC curve indicated. **(d)** Modulation across training of DMS D1⁺ early action-initiation excited neurons (*N* = 77 neurons/4 mice; average 19.25 neurons/mouse, s.e.m. = 5.36). Modulation index averaged across 0.5-s bins around action initiation. 2-way ANCOVA, Training: $F_{1.77, 133.28} = 1.24$, $P = 0.29$; Time bin: $F_{1.79, 133.96} = 13.96$, $P < 0.001$; Training x Time: $F_{3.40, 254.94} = 2.94$, $P = 0.028$. **(e)** Modulation across training of D1⁺ overtrain action-initiation excited neurons (*N* = 76 neurons/4 mice; average 19 neurons/mouse, s.e.m. 5.49). Modulation index around action initiation. 2-way ANCOVA, Training: $F_{1.8, 79.90} = 7.18$, $P = 0.002$; Time bin: $F_{1.64, 65.47} = 1.50$, $P = 0.23$; Training x Time: $F_{4.28, 171.34} = 1.50$, $P = 0.20$. Data presented as mean ± s.e.m. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

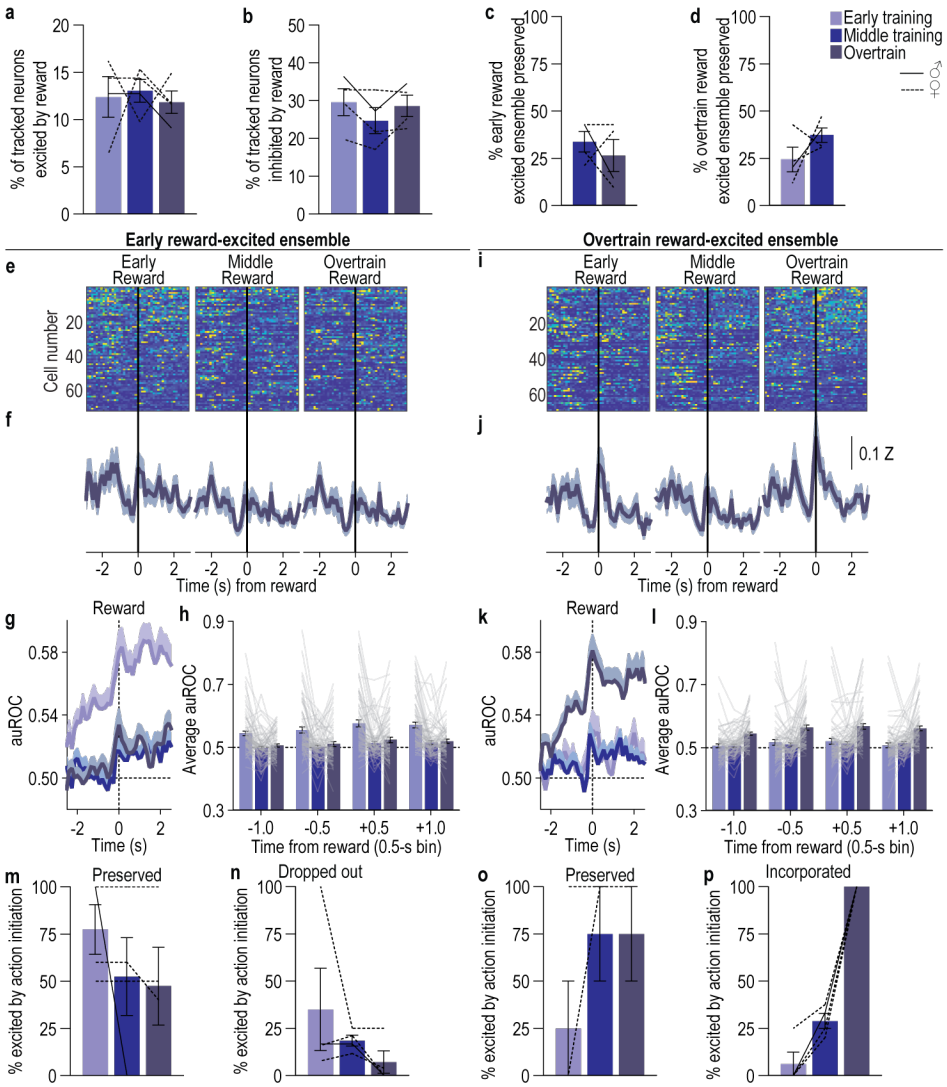


Extended Data Figure 2-3. An ensemble of DMS D1⁺ neurons is inhibited during action initiation across learning and as habits form. (a) Percent of all recorded coregistered DMS D1⁺ neurons (Average 137 coregistered neurons/mouse, s.e.m. 37.73) significantly inhibited by action initiation. Approximately 30 - 32% of DMS D1⁺ neurons were inhibited around action initiation and this did not change with training. 1-way ANOVA, $F_{1.29, 3.86} = 0.14$, $P = 0.79$. **(b-c)** Accuracy with which lever-press rate can be can be decoded from the activity of DMS D1⁺ early (b; 2-way ANOVA, Neuron activity: $F_{1, 3} = 53.88$, $P = 0.005$; Training: $F_{1.76, 5.29} = 0.09$, $P = 0.89$; Neuron activity x Training: $F_{1.73, 5.38} = 0.11$, $P = 0.88$) or overtrain (c; 2-way ANOVA, Neuron activity: $F_{1, 3} = 84.88$, $P = 0.003$; Training: $F_{1, 3} = 1.25$, $P = 0.35$; Neuron activity x Training: $F_{1, 3} = 1.30$, $P = 0.34$) action-initiation inhibited neurons. Lever-press rate can be reliably decoded from the activity of the population of DMS D1⁺ neurons that are inhibited by action initiation. **(d)** Percent of D1⁺ early action-initiation inhibited neurons that continued to be significantly inhibited by action initiation on the 4th and 8th training sessions. Approximately 52 - 55% of the early action-initiation inhibited ensemble continued to be inhibited around action initiation during the middle and overtraining phases of training. The proportion preserved did not change with training. 2-tailed t-test, $t_3 = 0.56$, $P = 0.62$, 95% CI -19.03 - 13.37. **(e)** Percent of D1⁺ overtrain action-initiation inhibited neurons that were also significantly inhibited by action initiation on 1st and 4th training sessions. Approximately 51 - 59% of the overtraining action-initiation inhibited ensemble was also inhibited around action initiation during the preceding early and middle training phases. The proportion preserved did not change with training. 2-tailed t-test, $t_3 = 2.02$, $P = 0.14$, 95% CI -4.52 - 20.15. **(f-i)** Activity and modulation across training of DMS D1⁺ early action-initiation inhibited neurons. Heat map of minimum to maximum deconvolved activity (sorted by total activity) (f), Z-scored activity (g), and area under the receiver operating characteristic curve (auROC) modulation index (h) of these cells around action initiation across training. (i) auROC modulation index averaged across 0.5-s bins around action initiation. 2-way ANCOVA, Training: $F_{1.65, 246.27} = 1.46$, $P = 0.24$; Time bin: $F_{2.16, 321.48} = 2.80$, $P = 0.06$; Training x Time: $F_{4.70, 699.66} = 0.27$, $P = 0.92$. The modulation of the early action-initiation inhibited ensemble did not significantly change with training. **(j-m)** Activity and modulation across training of DMS D1⁺ overtrain action-initiation inhibited neurons. Heat map (h), Z-scored activity (i), and auROC modulation index (j) of these cells around action initiation across training. (m) Modulation index around action initiation. 2-way ANCOVA, Training: $F_{1.82, 347.95} = 7.25$, $P = 0.001$; Time bin: $F_{1.91, 365.08} = 2.37$, $P = 0.10$; Training x Time: $F_{4.35, 831.34} = 0.67$, $P = 0.62$. This overtrain action-initiation inhibited ensemble became more inhibited by action initiation as training progressed. The D1⁺ neuron ensemble that was inhibited around action initiation was stable, similar to

the $D1^+$ neurons that were excited by action initiation. D1-cre: *N* = 4 (1 male). Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines.

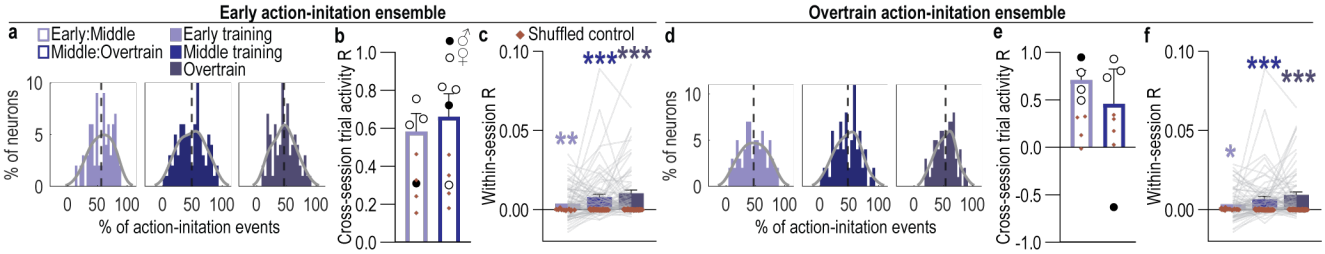


Extended Data Figure 2-4. A subensemble of DMS D1⁺ neurons is modulated by action termination during learning and as habits form. (a-b) Percent of all recorded coregistered DMS D1⁺ neurons (Average 137 coregistered neurons/mouse, s.e.m. 37.73) significantly excited (a; 1-way ANOVA, $F_{1.62, 4.85} = 0.37$, $P = 0.67$) or inhibited (b; 1-way ANOVA, $F_{1.17, 3.50} = 1.29$, $P = 0.34$) by action termination. Approximately 8 - 10% of coregistered D1⁺ neurons were excited by action termination, and 21 - 29% were inhibited. This did not change with training. **(c)** Percent of DMS D1⁺ early action-termination excited neurons ($N = 53$ neurons/4 mice; average 13.25 neurons/mouse, s.e.m. 5.73) that were then also significantly excited by action termination on the 4th (middle) and 8th (overtrain) training sessions. Approximately 30 - 37% of the early action-termination ensemble continued to be excited by action termination across training. The proportion preserved did not change with training. 2-tailed t-test, $t_3 = 0.85$, $P = 0.46$, 95% CI -20.34 - 35.07. **(d)** Percent of DMS D1⁺ overtrain action-termination excited neurons ($N = 58$ neurons/4 mice; average 14.5 neurons/mouse, s.e.m. 3.59) that were also significantly excited by action termination on the 1st (early) and middle training sessions. Approximately 59 - 76% of the overtraining action-termination ensemble was also excited by action termination during the preceding training phases. The proportion preserved did not significantly change with training. 2-tailed t-test, $t_3 = 2.52$, $P = 0.09$, 95% CI -4.48 to 38.31. **(e-h)** Activity and modulation across training of DMS D1⁺ early action-termination excited neurons. Heat map of minimum to maximum deconvolved activity (sorted by total activity) (d), Z-scored activity (e), and area under the receiver operating characteristic curve (auROC) modulation index (f) of these neurons around action termination across training. (h) auROC modulation index averaged across 0.5-s bins around action termination. 2-way ANCOVA, Training: $F_{1.92, 97.81} = 1.08$, $P = 0.34$; Time bin: $F_{1.65, 84.19} = 0.48$, $P = 0.59$; Training x Time: $F_{4.20, 214.10} = 1.33$, $P = 0.26$. Modulation of this early action-termination ensemble did not significantly change with training. **(i-l)** Activity and modulation across training of DMS D1⁺ overtrain action-termination excited neurons. Heat map (h), Z-scored activity (i), and auROC modulation index (j) of these cells around action termination across training. (l) Modulation index around action termination. 2-way ANOVA, Training: $F_{1.77, 99.05} = 1.14$, $P = 0.32$; Time bin: $F_{1.78, 99.88} = 0.03$, $P = 0.96$; Training x Time: $F_{4.20, 235.08} = 0.71$, $P = 0.60$. Modulation of the overtrain action-termination ensemble did not significantly change as training progressed. D1-cre: $N = 4$ (1 male). Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines.

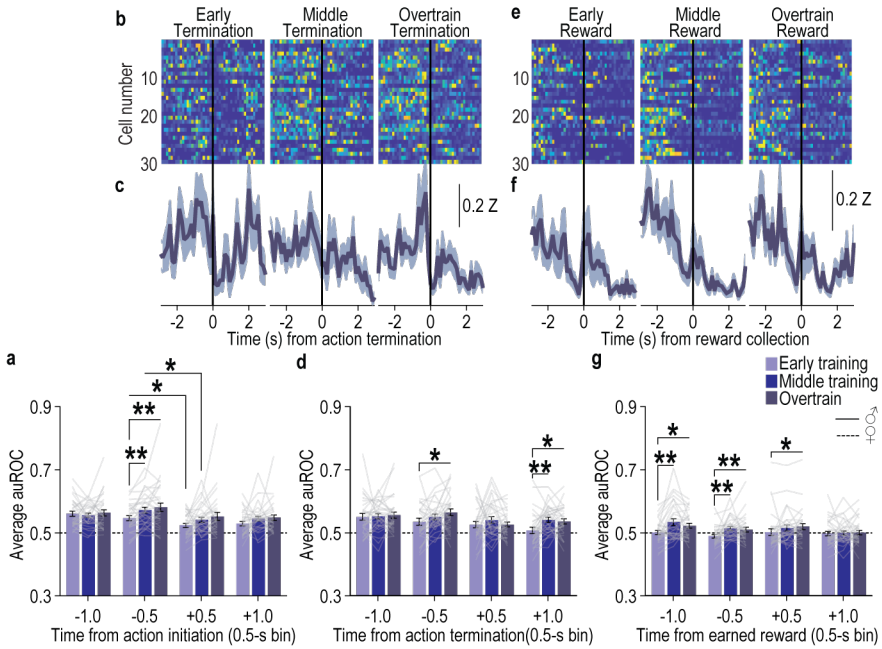


Extended Data Figure 2-5. An ensemble of DMS D1⁺ neurons is modulated by reward during learning and as habits form. (a-b) Percent of all recorded coregistered DMS D1⁺ neurons (Average 137 coregistered neurons/mouse, s.e.m. 37.73) significantly excited (a; 1-way ANOVA, $F_{1.46, 4.38} = 0.12$, $P = 0.83$) or inhibited (b; 1-way ANOVA, $F_{1.91, 5.76} = 2.52$, $P = 0.16$) by earned reward. Approximately 12 - 13% of coregistered D1⁺ neurons were excited by reward and 25 - 30% were inhibited. This did not change with training. **(c)** Percent of D1⁺ early reward excited neurons (N = 73 neurons/4 mice; average 18.25 neurons/mouse, s.e.m. 7.36) that continued to be significantly excited by reward during the middle and overtraining phases. Approximately 26 - 34% of the early reward ensemble continued to be excited by reward across training. The proportion preserved did not change with training. 2-tailed t-test, $t_3 = 0.70$, P =0.53, 95% CI -40.49 - 25.89. **(d)** Percent of DMS D1⁺ overtrain reward excited neurons (N = 70 neurons/4 mice; average 17.5 neurons/mouse, s.e.m. 6.34) that were significantly excited by reward on the early and middle training sessions. Approximately 24 - 37% of the overtraining reward ensemble was also excited by reward during the preceding training phases. The proportion preserved did not change with training. 2-tailed t-test, $t_3 = 1.31$, P =0.28, 95% CI -18.55 - 44.33. **(e-h)** Activity and modulation across training of DMS D1⁺ early reward excited neurons. Heat map (e), Z-scored activity (f), and auROC modulation index (g) of these cells around reward across training. (h) Modulation index around reward. 2-way ANCOVA, Training: $F_{1.74, 123.26} = 0.82$, $P = 0.42$; Time bin: $F_{2.45, 175.73} = 6.07$, $P = 0.01$; Training x Time: $F_{4.82, 342.52} = 1.59$, $P = 0.16$. This early reward ensemble was more excited after earned reward than before, suggesting modulation by reward experience. This did not significantly change as training progressed. **(i-l)** Activity and modulation across training of D1⁺ overtrain reward excited neurons. Heat map (i), Z-scored activity (j), and auROC modulation index (k) of these cells around reward on across training. (l) Modulation index around reward. 2-way ANCOVA Training: $F_{1.77, 120.87} = 2.26$, $P = 0.12$; Time bin: $F_{2.36, 160.42} = 1.30$, $P = 0.28$; Training x Time: $F_{5.37, 365.26} = 0.28$, $P = 0.93$. Modulation of the overtrain reward ensemble did not significantly change with training. **(m)** Percentage of the preserved early reward-excited neurons (those that continued to encode reward

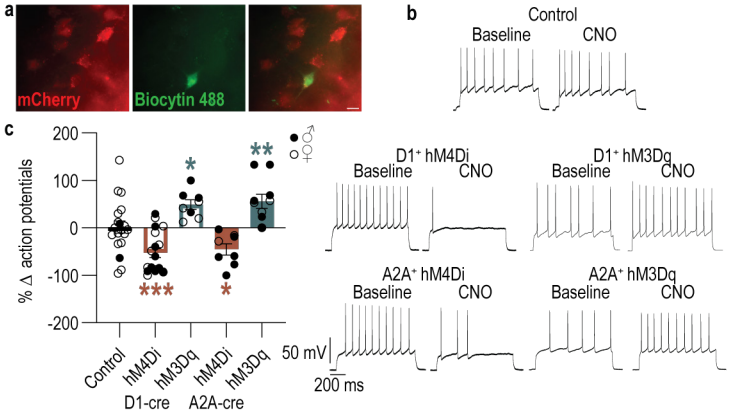
across training) that were also significantly excited by action initiation. 1-way ANOVA, $F_{1.06, 3.18} = 1.27$, $P = 0.34$. A substantial proportion of the neurons that continue to encode reward across training also encode action initiation. **(n)** Percentage of the neurons that dropped out of the early reward-excited ensemble (those that stopped being modulated by reward with training) that were significantly excited by action initiation. 1-way ANOVA, $F_{1.04, 3.13} = 1.78$, $P = 0.27$. Neurons that dropped out of the reward ensemble tended to be less likely to also encode action initiation. **(o)** Percentage of the preserved overtrain reward-excited neurons (those that encoded reward across training) that were also significantly excited by action initiation. 1-way ANOVA, $F_{1.00, 3.00} = 3.00$, $P = 0.18$. A substantial proportion of the neurons that encoded reward across training also encode action initiation. **(p)** Percentage of the neurons that were incorporated into the overtrain reward-excited ensemble (those that became modulated by reward with training) that were significantly excited by action initiation. 1-way ANOVA, $F_{1.48, 4.44} = 194.3$, $P < 0.0001$. Strikingly, all of the neurons that picked up encoding of the reward with training also encoded action initiation. Those neurons that either stably encode or pick up encoding of reward are also those that encode action initiation. Whereas, those neurons that stop encoding the reward were less likely to encode action initiation. Consistent with the emergent similarity of the population representation of action and outcome, this data suggests that with training individual neurons encode both the instrumental action and its reward outcome. D1-cre: $N = 4$ (1 male). Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines.



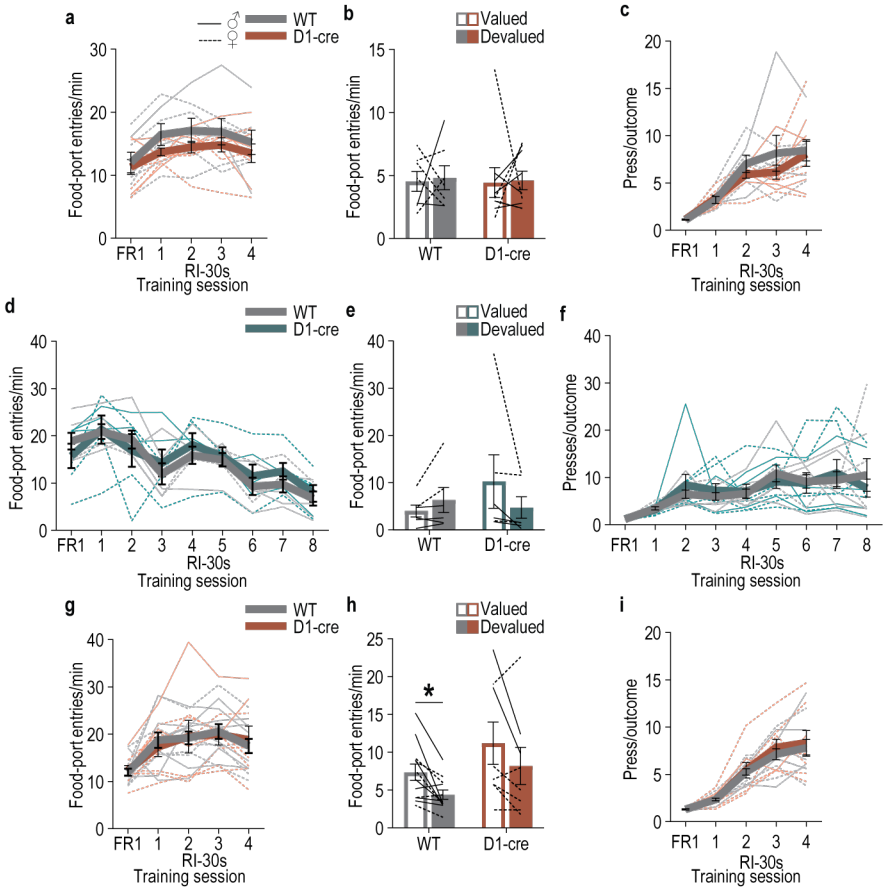
Extended Data Figure 2-6. DMS D1⁺ neurons encode action initiation with high fidelity. (a-c) Fidelity with which D1⁺ early action-initiation excited neurons encode action initiation. **(a)** Distribution of the percentage of early action-initiation excited neurons as a function of the percentage of action-initiation events to which they respond for each training phase. **(b)** Cross-session correlation of the response distributions. 2-way ANOVA, Neuron activity distribution (v. shuffled): $F_{1, 3} = 16.06$, $P = 0.03$; Training session: $F_{1, 3} = 0.25$, $P = 0.65$; Distribution x Training: $F_{1, 3} = 0.11$, $P = 0.76$. Early action-initiation D1⁺ neurons tend to respond on more than half the action-initiation events across training and this is consistent across training. **(c)** Within-session correlation of the activity around action initiation of each early action-initiation excited neuron. 2-way ANCOVA, Neuron activity: $F_{1, 74} = 27.35$, $P < 0.001$; Training: $F_{2, 148} = 5.84$, $P = 0.004$; Activity x Time: $F_{2, 148} = 6.28$, $P = 0.002$. The activity of early action-initiation excited D1⁺ neurons around action initiation is correlated above shuffled control within a training session and becomes more correlated with training. **(d-e)** Fidelity with which D1⁺ overtrain action-initiation excited neurons encode action initiation. **(d)** Distribution of the percentage of overtrain action-initiation excited neurons as a function of the percentage of action-initiation events to which they respond for each training phase. **(e)** Cross-session correlation of the response distributions. 2-way ANOVA, Neuron activity distribution: $F_{1, 3} = 19.36$, $P = 0.02$; Training: $F_{1, 3} = 0.27$, $P = 0.64$; Distribution x Training: $F_{1, 3} = 0.35$, $P = 0.60$. Overtrain action-initiation D1⁺ neurons tend to respond on more than half the action-initiation events across training and this is consistent across training. **(f)** Within-session correlation of the activity around action initiation of each overtrain action-initiation excited neuron. 2-way ANCOVA, Neuron activity: $F_{1, 73} = 19.32$, $P < 0.001$; Training: $F_{2, 146} = 1.87$, $P = 0.16$; Activity x Time: $F_{2, 146} = 1.96$, $P = 0.15$. The activity of overtrain action-initiation excited D1⁺ neurons around action initiation is correlated above shuffled control within a training session and becomes more correlated with training. D1-cre: $N = 4$ (1 male). Data presented as mean ± s.e.m. Males = closed circles, Females = open circles.



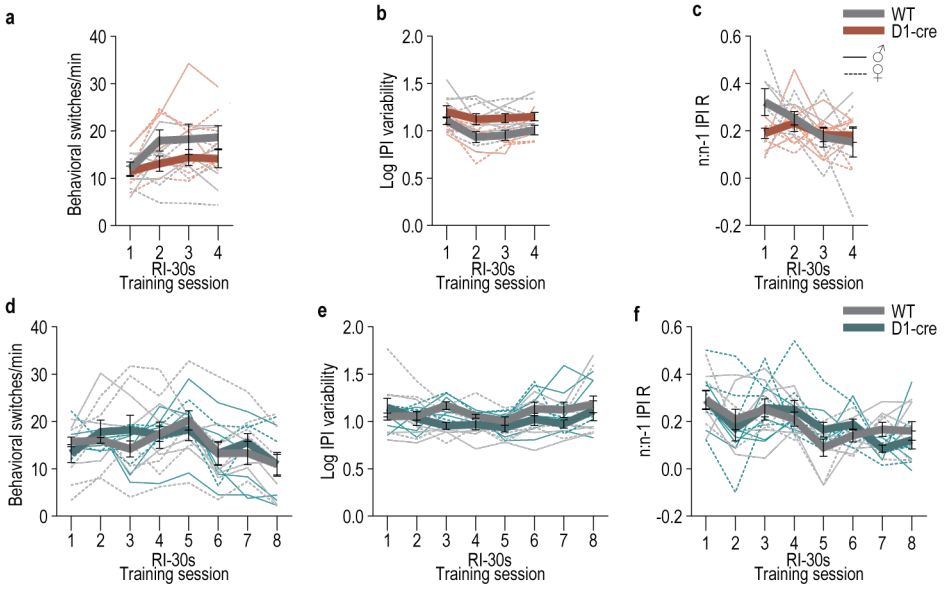
Extended Data Figure 2-7. A subensemble of stable DMS D1⁺ action-initiation excited neurons develop encoding of action termination and earned reward with training. We identified an ensemble (*N* = 31 neurons/4 D1-cre mice, 1 male; average 7.75 neurons/mouse, s.e.m. = 2.56) of D1⁺ neurons that stably encoded action-initiation across all phases of training. **(a)** Modulation, averaged in 0.5-s bins around action initiation of these neurons. 2-way ANCOVA, Training: $F_{1.70, 49.21} = 0.04$, $P = 0.94$; Time bin: $F_{1.73, 50.27} = 5.02$, $P = 0.01$; Training x Time: $F_{2.28, 123.97} = 2.78$, $P = 0.03$. Stable action-initiation excited D1⁺ neurons are more activated prior to action initiation than after and this modulation improves with training. **(b-c)** Heat map of minimum to maximum deconvolved activity (sorted by total activity) (b) and Z-scored (c) activity around action termination across training of the D1⁺ stable action-initiation excited neurons. **(d)** Modulation, averaged in 0.5-s bins around action termination of these neurons. 2-way ANCOVA, Training: $F_{1.87, 54.10} = 1.01$, $P = 0.37$; Time bin: $F_{2.00, 57.95} = 1.48$, $P = 0.23$; Training x Time: $F_{3.62, 104.97} = 1.98$, $P = 0.11$. **(e-f)** Heat map (e) and Z-scored (f) activity around reward collection across training of D1⁺ stable action-initiation excited neurons. **(g)** Modulation, averaged in 0.5-s bins around earned reward of these neurons. 2-way ANCOVA, Training: $F_{1.99, 57.57} = 2.61$, $P = 0.08$; Time bin: $F_{1.84, 53.36} = 0.80$, $P = 0.45$; Training x Time: $F_{3.87, 112.10} = 1.74$, $P = 0.15$. Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines.



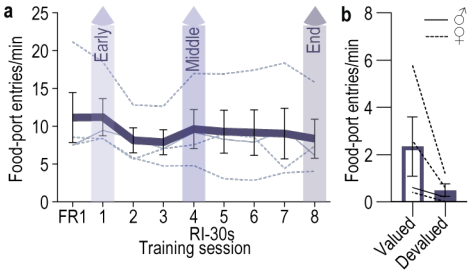
Extended Data Figure 3-1: Validation of chemogenetic approach to inhibit or activate DMS D1⁺ or A2A⁺ neurons. Fluorescence-guided, whole-cell patch clamp recordings in current-clamp mode were used to validate the efficacy of chemogenetic manipulation of DMS D1⁺ and A2A⁺ neurons. **(a)** Representative immunofluorescent image of a biocytin-filled, mCherry-positive cell. **(b)** Representative recordings of action potentials in single cells before (Baseline) and after CNO (10 μM for hM3Dq and 100 μM for hM4Di) bath application. Current injection intensity applied to induce action potential firing was 200 (Control, D1⁺ hM3Dq, A2A⁺ hM3Dq) or 250 pA (all others), with the same intensity used for baseline and CNO recordings. Cell membrane potentials were held at -80 mV using negative current to ensure consistent baseline conditions across recordings. **(c)** Percent change in action potentials under CNO, relative to pre-CNO baseline. 1-way ANOVA, $F_{4, 58} = 13.29$, $P < 0.0001$. Control, $N = 19$ cells, 4 D1-cre mice (1 male); D1-cre, hm4Di: $N = 8$ cells, 4 mice (3 males), hM3Dq: $N = 10$ cells, 5 mice (3 males); A2A-cre, hM4Di: $N = 8$ cells, 4 mice (3 males), hM3Dq: $N = 10$ cells, 5 mice (3 males). Data presented as mean ± s.e.m. Males = closed circles, Females = open circles. Scale bar = 10 μm. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, uncorrected. We were able to effectively chemogenetically inhibit and activate action potentials in both D1⁺ and A2A⁺ DMS neurons.



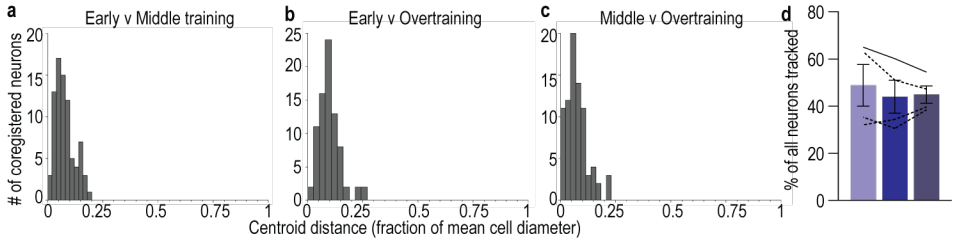
Extended Data Figure 3-2. Food-port entries during training and test with DMS D1⁺ neuron chemogenetic manipulation. (a-c) Chemogenetic inactivation of DMS D1⁺ neurons during learning. WT: N = 7 (2 males); D1-cre: N = 9 (5 males). **(a)** Training entry rate. 2-way ANOVA, Training: $F_{2.35, 32.91} = 7.40$, $P = 0.001$; Genotype: $F_{1, 14} = 1.30$, $P = 0.27$; Training x Genotype: $F_{4, 56} = 0.46$, $P = 0.77$. **(b)** Test entry rate. 2-way ANOVA, Value x Genotype: $F_{1, 14} = 0.003$, $P = 0.96$; Value: $F_{1, 14} = 0.05$, $P = 0.83$; Genotype: $F_{1, 14} = 0.03$, $P = 0.86$. **(c)** Training average presses/earned reward outcome. 2-way ANOVA, Training: $F_{2.05, 28.68} = 32.52$, $P < 0.0001$; Genotype: $F_{1, 14} = 0.57$, $P = 0.46$; Training x Genotype: $F_{4, 56} = 0.77$, $P = 0.55$. **(d-f)** Chemogenetic activation of DMS D1⁺ neurons during learning. WT: N = 6 (4 males); D1-cre: N = 6 (3 males). **(d)** Training entry rate. 2-way ANOVA, Training: $F_{2.38, 23.79} = 12.75$, $P < 0.0001$; Genotype: $F_{1, 10} = 0.21$, $P = 0.74$; Training x Genotype: $F_{8, 80} = 0.74$, $P = 0.65$. **(d)** Test entry rate. 2-way ANOVA, Value x Genotype: $F_{1, 10} < 0.0001$, $P > 0.99$; Value: $F_{1, 10} = 0.89$, $P = 0.37$; Genotype: $F_{1, 10} = 0.51$, $P = 0.49$. **(f)** Training average presses/earned reward outcome. 2-way ANOVA, Training: $F_{4.03, 60.47} = 9.42$, $P < 0.0001$; Genotype: $F_{1, 15} = 0.0001$, $P = 0.99$; Training x Genotype: $F_{8, 120} = 0.60$, $P = 0.78$. **(g-i)** Chemogenetic inactivation of DMS D1⁺ neurons at test of behavioral strategy after learning. WT: N = 12 (7 males); D1-cre: N = 12 (5 males). **(g)** Training entry rate. 2-way ANOVA, Training: $F_{3.01, 54.19} = 13.85$, $P < 0.0001$; Genotype: $F_{1, 18} < 0.0001$, $P > 0.99$; Training x Genotype: $F_{4, 72} = 0.34$, $P = 0.85$. **(h)** Test entry rate. 2-way ANOVA, Value x Genotype: $F_{1, 18} = 0.001$, $P = 0.97$; Value: $F_{1, 18} = 11.24$, $P = 0.004$; Genotype: $F_{1, 18} = 3.02$, $P = 0.10$. **(i)** Training average presses/earned reward outcome. 2-way ANOVA, Training: $F_{1.86, 44.59} = 96.52$, $P < 0.0001$; Genotype: $F_{2, 24} = 0.13$, $P = 0.88$; Training x Genotype: $F_{8, 96} = 0.13$, $P > 0.99$. Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines. *$P < 0.05$. Neither chemogenetic inhibition nor activation of DMS D1⁺ neurons affected checks of the food-delivery port or altered the press-reward action-outcome relationship.



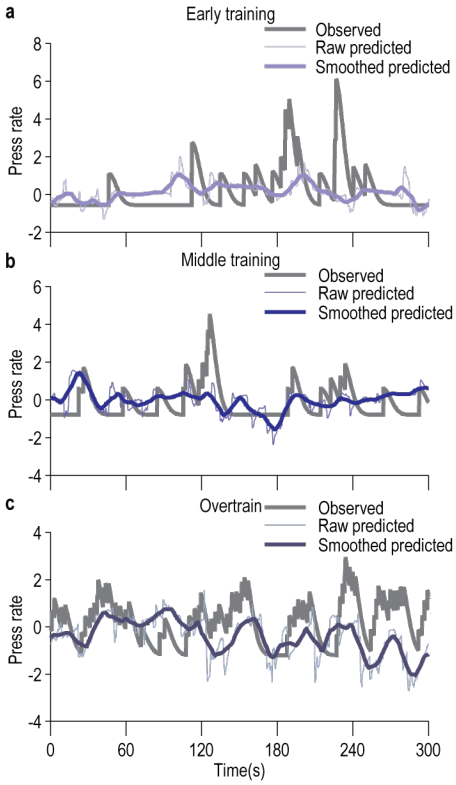
Extended Data Figure 3-3. Behavioral strategy variables during training with DMS D1⁺ neuron chemogenetic manipulation. (a-c) Chemogenetic inactivation of DMS D1⁺ neurons during learning. WT: *N* = 7 (2 males); D1-cre: *N* = 9 (5 males). Behavioral switch rate (a; how often subjects stop one behavior to transition to another i.e., press → food-port check or food-port check → press; 2-way ANOVA, Training: $F_{2.25, 31.49} = 5.74$, $P = 0.006$; Genotype: $F_{1, 14} = 2.25$, $P = 0.16$; Training x Genotype: $F_{2.25, 31.49} = 1.16$, $P = 0.33$), interpress interval variability (b; standard deviation of the log-transformed interpress-intervals; 2-way ANOVA, Genotype: $F_{1, 14} = 7.64$, $P = 0.02$; Training: $F_{2.38, 33.33} = 3.05$, $P = 0.05$; Training x Genotype: $F_{2.38, 33.33} = 0.43$, $P = 0.68$), and relationship to past trial behavior (c; correlation between the current interpress interval (n) and the interpress interval of the prior press (n-1); 2-way ANOVA, Genotype: $F_{1, 14} = 0.77$, $P = 0.39$; Training: $F_{2.62, 36.68} = 3.04$, $P = 0.05$; Training x Genotype: $F_{2.62, 36.68} = 1.93$, $P = 0.15$) across days of training. Consistent the evidence from the linear regression model of the variables associated with habit, inhibition of DMS D1⁺ neurons caused slightly lower behavioral switch rate and significantly lower consistency (higher interpress interval variability). It did not significantly affect how much current trial interpress wait time related to past trial wait times. **(d-f)** Chemogenetic activation of DMS D1⁺ neurons during learning. WT: *N* = 6 (4 males); D1-cre: *N* = 6 (3 males). Behavioral switch rate (d; 2-way ANOVA, Training: $F_{2.98, 44.64} = 5.53$, $P = 0.003$; Genotype: $F_{1, 15} = 0.03$, $P = 0.87$; Training x Genotype: $F_{2.98, 44.64} = 0.94$, $P = 0.43$), interpress interval variability (e; Training: $F_{2.21, 33.10} = 2.08$, $P = 0.13$; Genotype: $F_{1, 15} = 1.76$, $P = 0.21$; Training x Genotype: $F_{2.21, 33.10} = 1.43$, $P = 0.25$), and relationship to past trial behavior (f; 2-way ANOVA, Genotype: $F_{1, 15} = 0.003$, $P = 0.69$; Training: $F_{2.21, 33.10} = 2.08$, $P = 0.14$; Training x Genotype: $F_{2.21, 33.10} = 1.43$, $P = 0.25$) across days of training. Excitation of DMS D1⁺ neurons did not affect behavioral strategy variables. Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines.



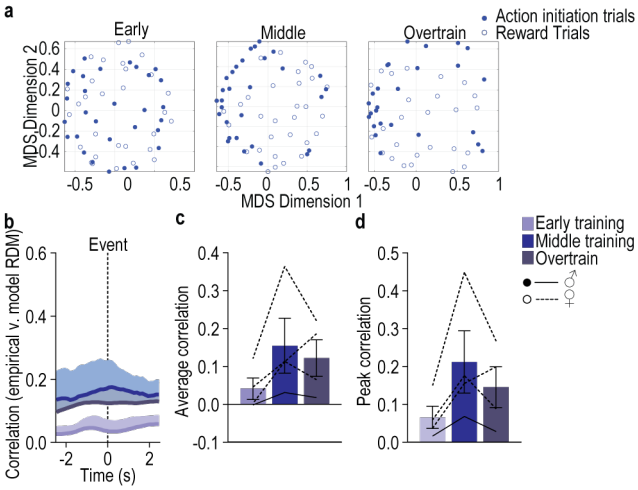
Extended Data Figure 4-1: Entries into the food-delivery port during training and test for DMS A2A$^+$ imaging experiment- habitual subjects. (a) Training entry rate. 1-way ANOVA, Training: $F_{2.69,\ 8.06} = 1.93$, $P = 0.20$ **(b)** Test entry rate. 2-tailed t-test, $t_3 = 1.87$, $P = 0.16$, 95% CI -5.00 - 1.30. Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines.



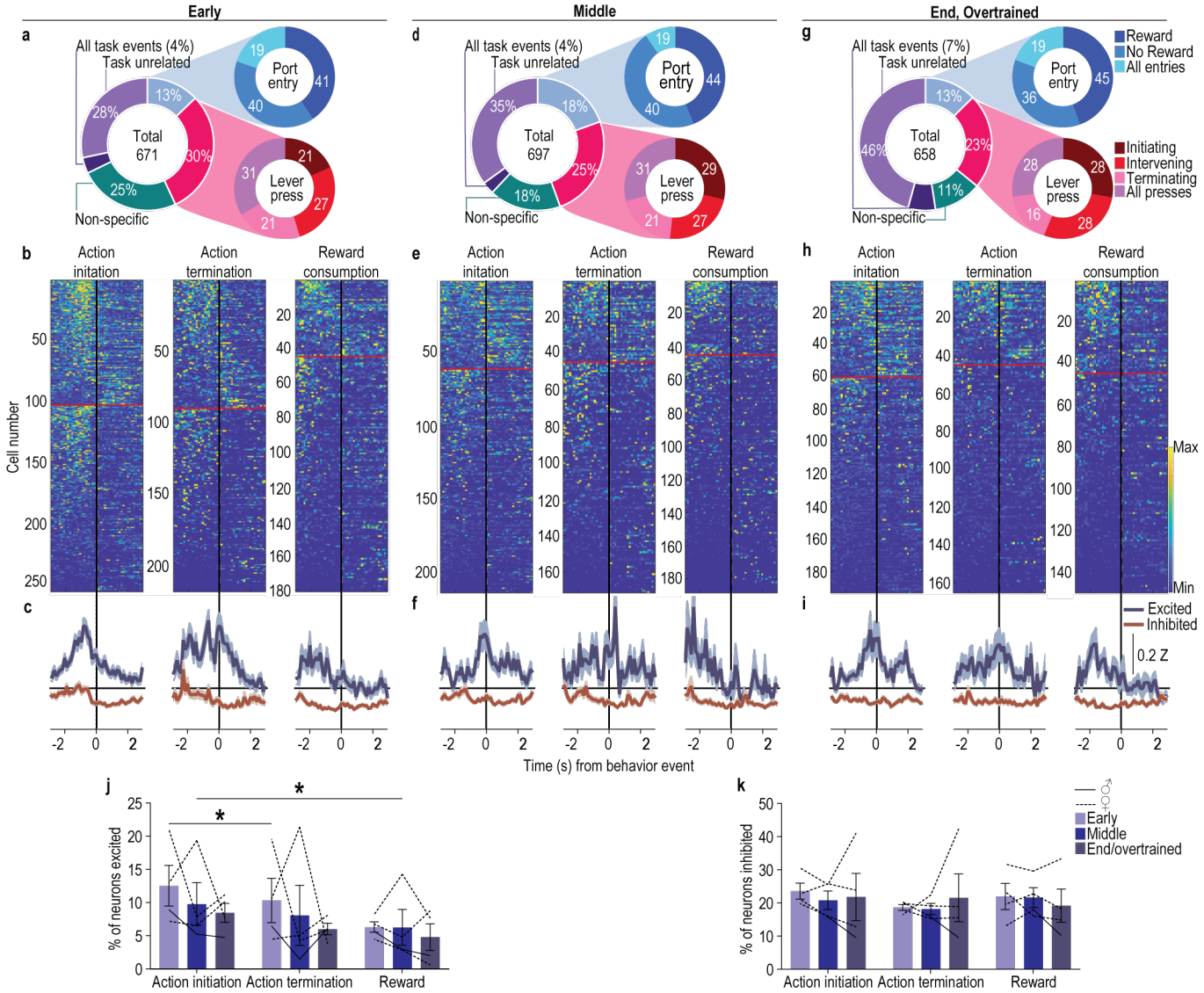
Extended Data Figure 4-2. Representative example of coregistration of DMS A2A⁺ neurons across training. (a) Distribution of the distance between cell centroids (as a fraction of total cell diameter) of co-registered cell pairs in the 1st (early) and 4th (middle) training sessions. **(b)** Distribution of centroid distance of co-registered cell pairs in the 1st and 8th (overtrain) training sessions. **(c)** Distribution of centroid distance of co-registered cell pairs in the 4th and 8th training sessions. **(d)** Percent of all neurons coregistered across training. 1-way ANOVA, $F_{1.18, 3.54} = 0.78$, $P = 0.45$. A2A-cre: $N = 4$ (1 male). We were able to coregister on average 47% (s.e.m. 3.60%) of the A2A⁺ neurons across all three training phases. Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines.



Extended Data Figure 4-3. Representative decoded press rate from the activity of coregistered DMS A2A^{+} neurons across training. (a-c) Actual observed and predicted lever-press rate from the 1st (a, early), 4th (b, middle), and 8th (c, overtrain) training session for a single subject.

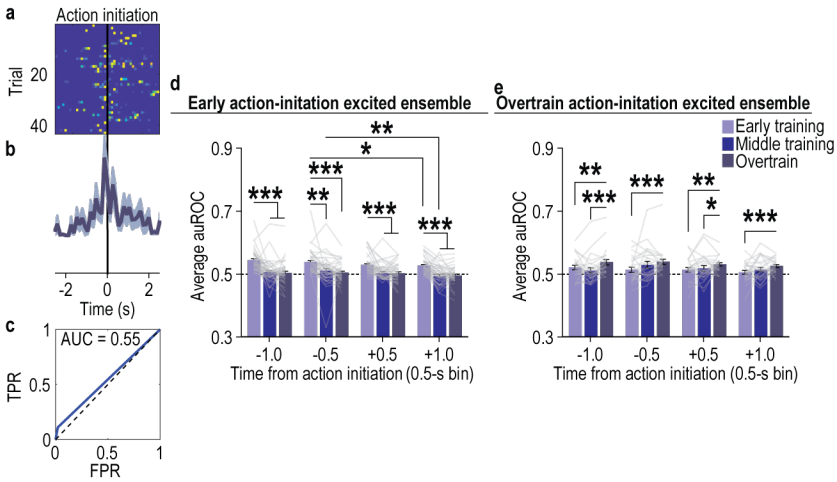


Extended Data Figure 4-4. DMS A2A⁺ population representation of action and reward does not become more similar with learning. We used a time-resolved Representational Similarity analysis to assess how the neural activity patterns for action initiation and reward evolve over the course of learning. **(a)** Representative single-subject example multidimensional scaling (MDS) plots of the Representational Dissimilarity Matrix (RDM) of pairwise dissimilarities between the DMS A2A⁺ population activity patterns around action initiation and earned reward from the 1st (a, early), 4th (b, middle), and 8th (c, overtrain) training session. **(b-d)** Spearman's rank correlation values between the empirical RDM of the A2A⁺ population action initiation and reward activity patterns and those of a theoretical RDM generated to hypothesize complete dissimilarity between action initiation and reward (lower correlation indicates similar neural representations between action initiation and reward) across time centered around each event (b), average (c; 1-way RM ANOVA, $F_{1.93, 5.08} = 3.62$, $P = 0.10$), or peak (d; 1-way RM ANOVA, $F_{1.69, 5.07} = 5.06$, $P = 0.06$) +/- 2.5 s around the events. With learning and continued training the DMS A2A⁺ neuronal population representation of action initiation and earned reward does not significantly change, but trends towards becoming less similar. A2A-cre: $N = 4$ (1 male). Data presented as mean ± s.e.m. Males = closed circles/solid lines, Females = open circles/dashed lines.

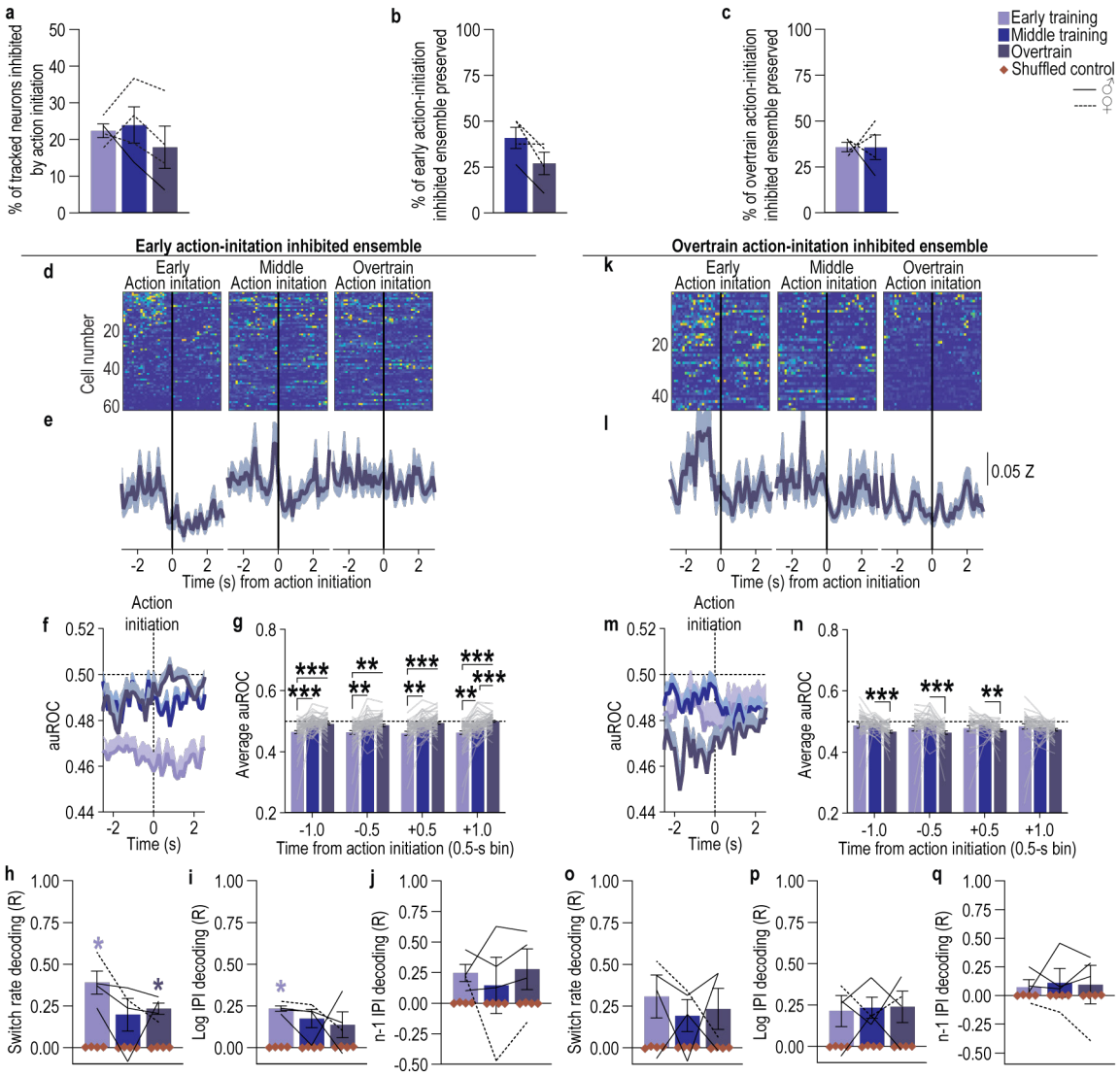


Extended data Figure 5-1. DMS A2A⁺ neurons are modulated by actions during instrumental learning and overtraining. To ask how DMS A2A⁺ neurons encode task events during action-outcome learning and the transition to habit, we computed auROC values and classified neurons as significantly modulated by critical behavioral events: lever-press action initiation (first press in the session, after earned reward collection, or after a non-reinforced food-port check), action termination (last press before reward collection or non-reinforced food-port check), intervening lever presses (all other presses), non-reinforced food-port checks, and reward collection. **(a-c)** Activity of DMS A2A⁺ neurons on the 1st (Early) session of RI training. (a) Percent of all recorded neurons (*N* = 671) significantly modulated by lever presses, food-delivery port checks, and reward. (b) Heat map of minimum to maximum deconvolved activity (sorted by total activity) of each DMS A2A⁺ neuron significantly modulated around lever-press action initiation (right), action termination (middle), or reward consumption (left). Above red line = excited, below = inhibited. (c) Z-scored activity of each population of modulated neurons. **(d-f)** Activity of DMS A2A⁺ neurons on the 4th (Middle) training session. (d) Percent of all recorded neurons (*N* = 697) significantly modulated by lever presses, food-delivery port checks, and reward. (e) Heat map of minimum to maximum deconvolved activity of each DMS A2A⁺ neuron significantly modulated around lever-press action initiation, action termination, or reward. (f) Z-scored activity of each population of modulated neurons. **(g-i)** Activity of DMS A2A⁺ neurons on the 8th (End, Overtrain) training session. (g) Percent of all recorded neurons (*N* = 658) significantly modulated by lever presses, food-delivery port checks, and reward. (h) Heat map of minimum to maximum deconvolved activity of each DMS A2A⁺ neuron significantly modulated around lever-press action initiation, action termination, or reward. (i) Z-scored activity of each population of modulated neurons.
(j-k) Percent of all recorded neurons (Early: *N* = 671, Middle: *N* = 697; End *N* = 658) classified (auROC values >95th percentile of the distribution of shuffled auROCs within 2 s before or after event) as significantly excited (j;

2-way ANOVA, Event: $F_{1.12, 3.36} = 23.83$, $P = 0.01$; Training session: $F_{1.32, 3.97} = 0.05$, $P = 0.54$; Training x Event: $F_{1.86, 5.59} = 0.29$, $P = 0.77$) or inhibited (k; 2-way ANOVA, Event: $F_{1.90, 5.69} = 2.03$, $P = 0.23$; Training session: $F_{1.12, 3.36} = 0.05$, $P = 0.86$; Training x Event: $F_{1.45, 4.33} = 0.90$, $P = 0.44$) around initiating lever presses, terminating lever presses, or reward consumption. Action initiation excited more DMS A2A$^+$ neurons than termination or reward. Similar proportions of the DMS A2A$^+$ neurons were inhibited around each event type. Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines. *$P < 0.05$.

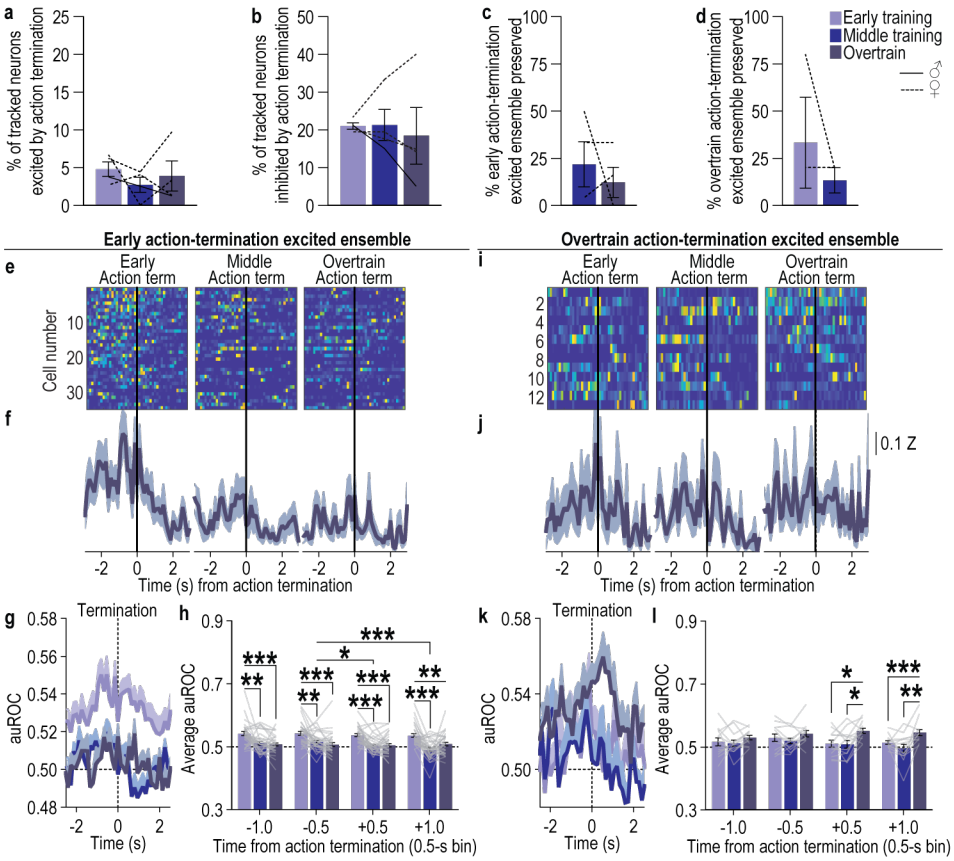


Extended Data Figure 5-2. Example and quantification of the modulation of DMS A2A⁺ action-initiation excited neurons. (a-c) Representative single-subject/single-session example of the trial x trial activity of a single action-initiation excited DMS A2A⁺ neuron. (a) Heat map of minimum to maximum deconvolved activity sorted by trial of a single A2A⁺ action-initiation excited neuron. (b) Z-scored, trial-averaged activity of this neuron. (c) ROC plot with true positive rate (TPR) and false positive rate (FPR) for this neuron compared to chance level discrimination (dashed line), with the area under the ROC curve indicated. **(d)** Modulation across training of DMS A2A⁺ early action-initiation excited neurons (N = 42 neurons/4 mice; average 16.8 neurons/mouse, s.e.m. = 5.61). Modulation index averaged across 0.5-s bins around action initiation. 2-way ANCOVA, Training: $F_{1.74, 69.82} = 4.10$, $P = 0.03$; Time bin: $F_{1.70, 68.14} = 0.18$, $P = 0.80$; Training x Time: $F_{4.70, 187.84} = 0.90$, $P = 0.68$. **(e)** Modulation across training of A2A⁺ overtrain action-initiation excited neurons (N = 25 neurons/4 mice; average 6.25 neurons/mouse, s.e.m. 2.69). Modulation index around action initiation. 2-way ANCOVA, Training: $F_{1.75, 40.22} = 1.15$, $P = 0.32$; Time bin: $F_{1.74, 39.90} = 0.10$, $P = 0.88$; Training x Time: $F_{4.23, 97.45} = 0.64$, $P = 0.70$. Data presented as mean ± s.e.m. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.



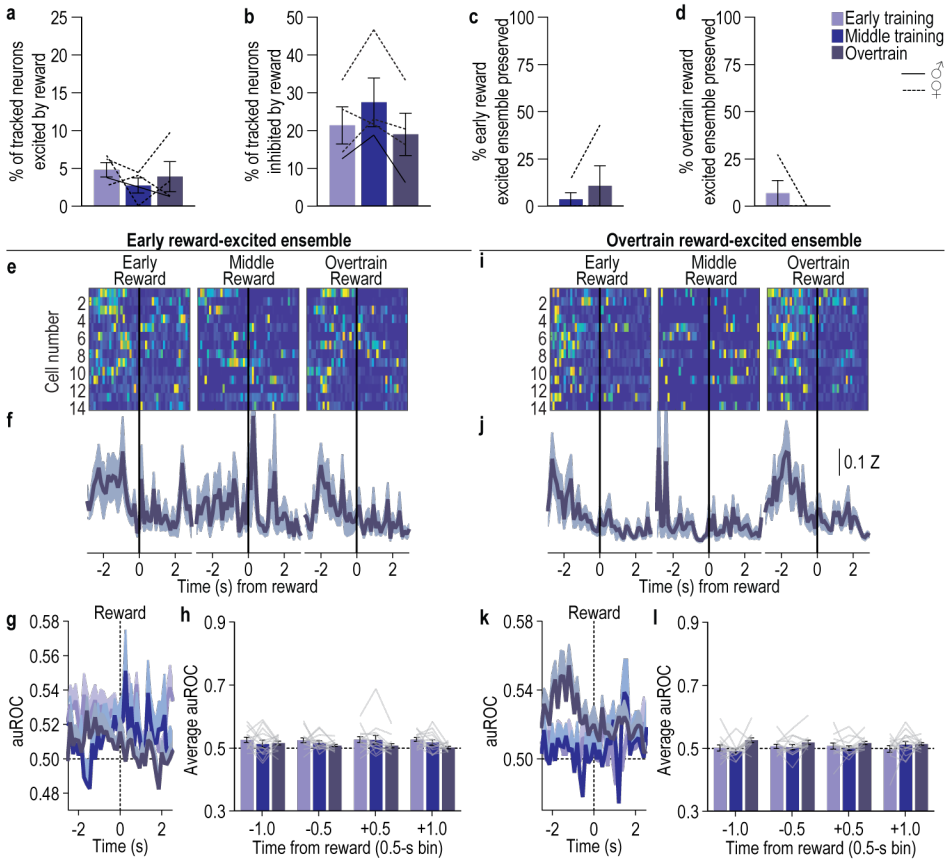
Extended Data Figure 5-3. The ensemble of DMS A2A⁺ neurons inhibited by action initiation shifts as habits form. (a) Percent of all recorded coregistered DMS A2A⁺ neurons (Average 73.25 coregistered neurons/mouse, s.e.m. 17.09) significantly inhibited by action initiation. Approximately 18 - 24% of coregistered DMS A2A⁺ neurons were inhibited by action initiation and this did not significantly change across training. 1-way ANOVA, $F_{1.05,\ 3.16} = 1.14$, $P = 0.37$. **(b)** Percent of DMS A2A⁺ early action-initiation-inhibited neurons that were then also significantly inhibited by action initiation on the 4th (middle) and 8th (overtrain) training sessions. Approximately 27 - 41% of the early action-initiation inhibited ensemble continued to be inhibited around action initiation during the middle and overtraining phases of training. The proportion preserved did not change with training. 2-tailed t-test, $t_3 = 2.70$, $P = 0.07$, 95% CI -30.41 - 2.52. **(c)** Percent of DMS A2A⁺ overtrain action-initiation-inhibited neurons that were significantly modulated by action initiation on prior the 1st (early) and 4th (middle) training sessions. Approximately 36% of the overtraining action-initiation inhibited ensemble was also inhibited around action initiation during the preceding early and middle training phases. The proportion preserved did not change with training. 2-tailed t-test, $t_3 = 0.01$, $P = 0.99$, 95% CI -29.03 - 28.80. **(d-g)** Activity and modulation across training of DMS A2A⁺ early action-initiation-inhibited neurons. Heat map of minimum to maximum deconvolved activity (sorted by total activity) (d), Z-scored activity (e), and area under the receiver operating characteristic curve (auROC) modulation index (f) of these cells around action initiation across training. **(g)** auROC modulation index averaged across 0.5-s bins around action initiation. 2-way ANCOVA, Training: $F_{1.38,\ 83.91} = 1.38$, $P = 0.26$; Time bin: $F_{2.48,\ 151.16} = 3.28$, $P = 0.02$; Training x Time: $F_{5.10,\ 311.30} = 3.80$, $P = 0.001$. This early action-initiation inhibited ensemble became less inhibited by action initiation as training progressed. **(h-j)** Linear decoding accuracy from DMS A2A⁺ early action-initiation inhibited neuron activity behavioral switch rate (h; 2-way ANOVA, Neuron activity: $F_{1,\ 3} = 33.71$, $P = 0.01$; Training: $F_{1.58,\ 4.74} = 2.39$, $P = 0.19$; Neuron activity x Training: $F_{1.61,\ 4.83} = 2.34$, $P = 0.19$), log interpress intervals (i; 2-way ANOVA, Neuron activity: $F_{1,\ 3} = 143.00$, $P =$

0.001; Training: $F_{1.03, 3.10} = 0.56$, $P = 0.51$; Neuron activity x Training: $F_{1.04, 3.12} = 0.55$, $P = 0.52$), and the prior press interpress interval (j; 2-way ANOVA, Neuron activity: $F_{1, 3} = 2.60$, $P = 0.21$; Training: $F_{1.03, 3.02} = 0.37$, $P = 0.59$; Neuron activity x Training: $F_{1.01, 3.02} = 0.34$, $P = 0.60$). Both behavioral switch rate and relative interpress interval, but not prior press interpress interval, could be decoded from the activity of the early action-initiation inhibited A2A$^+$ neurons. **(k-l)** Activity and modulation across training of coregistered DMS A2A$^+$ overtrain action-initiation-inhibited neurons. Heat map of minimum to maximum deconvolved activity (k), Z-scored activity (l), and area under the receiver operating characteristic curve (auROC) modulation index (m) of these cells around action initiation across training. **(m)** auROC modulation index averaged across 0.5-s bins around action initiation. 2-way ANCOVA, Training: $F_{1.21, 53.15} = 5.67$, $P = 0.005$; Time bin: $F_{2.62, 115.18} = 2.16$, $P = 0.10$; Training x Time: $F_{4.48, 196.92} = 3.56$, $P = 0.006$. This overtrain action-initiation inhibited ensemble became more inhibited prior to action initiation as training progressed. **(o-q)** Linear decoding accuracy from DMS A2A$^+$ overtrain action-initiation inhibited neuron activity of behavioral switch rate (o; 2-way ANOVA, Neuron activity: $F_{1, 3} = 10.48$, $P = 0.05$; Training: $F_{1.68, 5.04} = 0.28$, $P = 0.73$; Neuron activity x Training: $F_{1.63, 4.89} = 0.30$, $P = 0.71$), the variability of the interpress intervals (p; 2-way ANOVA, Neuron activity: $F_{1, 3} = 12.94$, $P = 0.04$; Training: $F_{1.23, 3.70} = 0.04$, $P = 0.89$; Neuron activity x Training: $F_{1.22, 3.65} = 0.03$, $P = 0.90$), and the prior press interpress interval (q; 2-way ANOVA, Neu0on activity: $F_{1, 3} = 0.80$, $P = 0.44$; Training: $F_{1.78, 5.34} = 0.04$, $P = 0.95$; Neuron activity x Training: $F_{1.81, 5.44} = 0.34$, $P = 0.95$). Both behavioral switch rate and relative interpress interval, but not prior press interpress interval, could be decoded from the activity of the overtrain action-initiation inhibited A2A$^+$ neurons. A2A-cre: $N = 4$ (1 male). Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

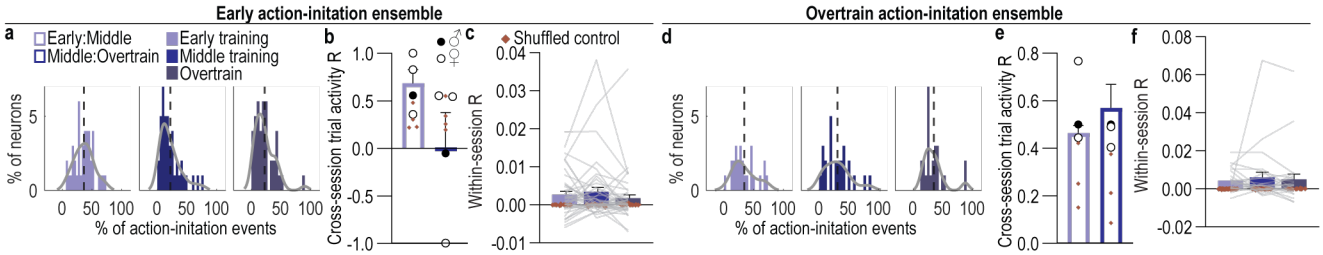


Extended Data Figure 5-4. The ensemble of DMS A2A⁺ neurons encoding action termination shifts as habits form. (a-b) Percent of all recorded coregistered DMS A2A⁺ neurons (Average 73.25 coregistered neurons/mouse, s.e.m. 17.09) significantly excited (a; 1-way ANOVA, $F_{1.48, 4.43} = 1.12$, $P = 0.38$) or inhibited (b; 1-way ANOVA, $F_{1.02, 3.05} = 0.20$, $P = 0.69$) by action termination. Approximately 4 - 10% of coregistered DMS A2A⁺ neurons were excited by action termination and 19 - 21% were inhibited. This did not change with training. $F_{1.00, 3.02} = 1.09$, $P = 0.37$. **(c)** Percent of DMS A2A⁺ early action-termination excited neurons ($N = 35$ neurons/4 mice; average 8.75 neurons/mouse, s.e.m. 5.45) that were also significantly excited by action termination on the 4th (middle) and 8th (overtrain) training sessions. Only 12 - 22% of the early action-termination ensemble continued to be excited by action termination during the middle and overtraining phases of training. The proportion preserved did not change with training. Middle v. Overtrain $t_3 = 0.69$, $P = 0.54$. **(d)** Percent of DMS A2A⁺ overtrain action-termination excited ($N = 13$ neurons/4 mice; average 3.25 neurons/mouse, s.e.m. 1.18) neurons that were also significantly excited by action termination on the 1st (early) and 4th (middle) training sessions. Only 13 - 33% of the overtraining action-termination ensemble was also excited by action termination during the preceding early and middle training phases. The proportion preserved did not change with training. Early v. Middle $t_3 = 1.00$, $P = 0.42$. **(e-h)** Activity and modulation across training of DMS A2A⁺ early action-termination excited neurons. Heat map of minimum to maximum deconvolved activity (sorted by total activity) (e), Z-scored activity (f), and area under the receiver operating characteristic curve (auROC) modulation index (g) of these cells around action termination across training. (h) auROC modulation index averaged across 0.5-s bins around action termination. Training: $F_{1.66, 54.75} = 3.77$, $P = 0.03$; Time bin: $F_{1.82, 60.19} = 1.17$, $P = 0.33$; Training x Time: $F_{4.19, 138.28} = 1.01$, $P = 0.41$. This early action-termination ensemble became less modulated by action termination as training progressed. **(i-l)** Activity and modulation across training of DMS A2A⁺ overtrain action-termination excited neurons. Heat map of minimum to maximum deconvolved activity (i), Z-scored activity (j), and area under the receiver operating characteristic curve (auROC) modulation index (j) of these cells around action termination across training. (k) auROC modulation index averaged across 0.5-s bins around action termination. Training: $F_{1.82, 19.98} = 3.20$, $P = 0.07$; Time bin: $F_{1.55, 17.10} = 2.28$, $P = 0.10$; Training x Time: $F_{3.25, 37.70} = 0.31$, $P = 0.83$. This overtrain action-termination ensemble became slightly more modulated by action

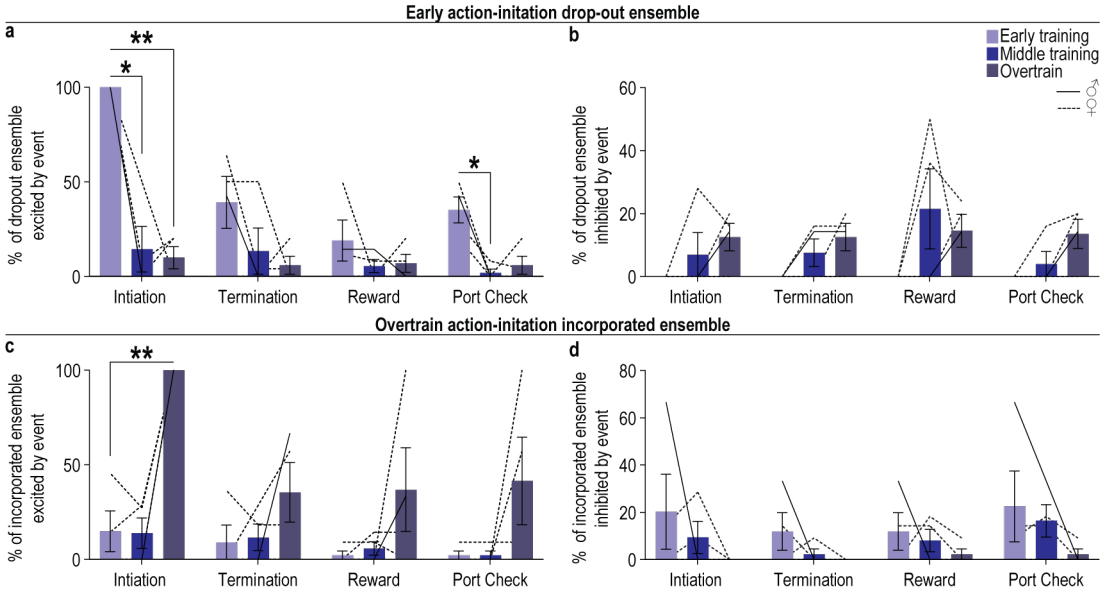
termination as training progressed. A2A-cre: $N = 4$ (1 male). Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.



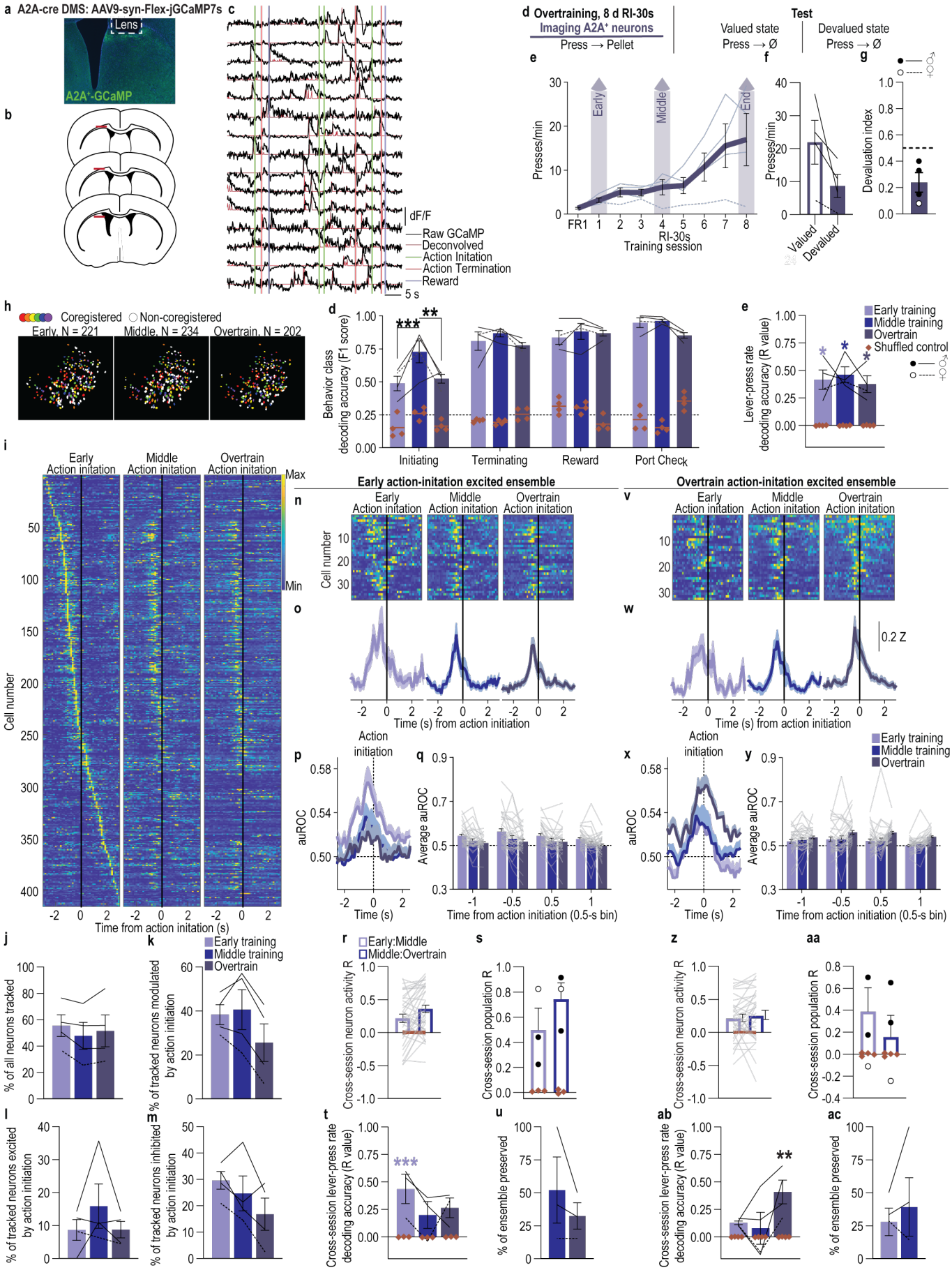
Extended Data Figure 5-5. Few DMS A2A⁺ neurons are excited by reward. (a-b) Percent of coregistered DMS A2A⁺ neurons significantly excited (a; 1-way ANOVA, $F_{1.89, 5.68} = 0.75$, $P = 0.51$) or inhibited (b; 1-way ANOVA, $F_{1.76, 5.28} = 3.52$, $P = 0.11$) by collection of the earned reward. Only 3 - 5% of coregistered DMS A2A⁺ neurons were excited by reward, substantially more (19 - 27%) were inhibited by reward. **(c)** Percent of DMS A2A⁺ early reward excited neurons ($N = 14$ neurons/4 mice; average 3.50 neurons/mouse, s.e.m. 1.19) that were then significantly excited by reward on the 4th (middle) and 8th (overtrain) training sessions. Only 4 - 11% of the small early reward ensemble continued to be modulated by reward during the middle and overtraining phases of training. The proportion preserved did not change with training. Middle v. Overtrain 2-tailed Wilcoxon signed rank test, W = 1.00, P > 0.99. **(d)** Percent of DMS A2A⁺ overtrain reward excited neurons ($N = 14$ neurons/4 mice; average 3.50 neurons/mouse, s.e.m. 2.50) that were significantly excited by reward on the 1st (early) and 4th (middle) training sessions. Only 0 - 7% of the small overtraining reward ensemble was also modulated by reward during the preceding early and middle training phases. The proportion preserved did not change with training. Early v. Middle 2-tailed Wilcoxon signed rank test, W = -1.00, P > 0.99. **(e-h)** Activity and modulation across training of DMS A2A⁺ early reward excited neurons. Heat map of minimum to maximum deconvolved activity (e), Z-scored activity (f), and area under the receiver operating characteristic curve (auROC) modulation index (g) of these cells around reward across training. (h) auROC modulation index averaged across 0.5-s bins around reward. Training: $F_{1.53, 18.36} = 0.492$, $P = 0.57$; Time bin: $F_{1.54, 18.49} = 0.64$, $P = 0.50$; Training x Time: $F_{2.58, 30.98} = 0.41$, $P = 0.71$. **(i-l)** Activity and modulation across training of DMS A2A⁺ overtrain reward excited neurons. Heat map of minimum to maximum deconvolved activity (i), Z-scored activity (j), and area under the receiver operating characteristic curve (auROC) modulation index (k) of these cells around reward on across training. (l) auROC modulation index averaged across 0.5-s bins around reward. Training: $F_{1.51, 18.10} = 1.07$, $P = 0.36$; Time bin: $F_{1.98, 23.76} = 2.78$, $P = 0.08$; Training x Time: $F_{2.94, 35.24} = 0.65$, $P = 0.58$. A2A-cre: $N = 4$ (1 male). Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines.



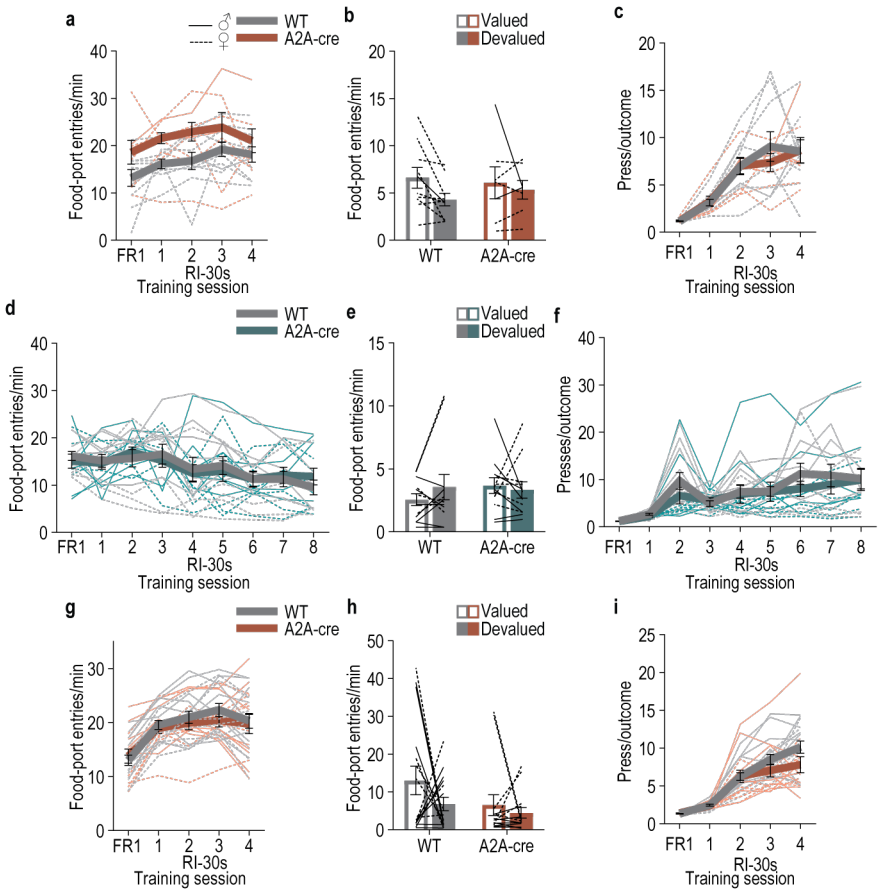
Extended Data Figure 5-6. Fidelity with which DMS A2A⁺ neurons encode action initiation. (a-c) Fidelity with which A2A⁺ early action-initiation excited neurons encode action initiation. **(a)** Distribution of the percentage of early action-initiation excited neurons as a function of the percentage of action-initiation events to which they respond for each training phase. **(b)** Cross-session correlation of the response distributions. 2-way ANOVA, Neuron activity distribution (v. shuffled): $F_{1, 3} = 0.03$, $P = 0.88$; Training sessions: $F_{1, 3} = 1.34$, $P = 0.33$; Distribution x Training: $F_{1, 3} = 2.99$, $P = 0.18$. Early action-initiation A2A⁺ neurons tend to respond on less than half of the action-initiation events and this decreases with training and is not correlated across training sessions. **(c)** Within-session correlation of the activity around action initiation of each early action-initiation excited neuron. 2-way ANCOVA, Neuron activity: $F_{1, 36} = 2.71$, $P = 0.11$; Training: $F_{2, 72} = 2.47$, $P = 0.09$; Activity x Time: $F_{2, 72} = 2.45$, $P = 0.09$. The activity of early action-initiation excited A2A⁺ neurons around action initiation is not significantly correlated within a training session. **(d-e)** Fidelity with which A2A⁺ overtrain action-initiation excited neurons encode action initiation. **(d)** Distribution of the percentage of overtrain action-initiation excited neurons as a function of the percentage of action-initiation events to which they respond for each training phase. **(e)** Cross-session correlation of the response distributions. 2-way ANOVA, Neuron activity distribution: $F_{1, 2} = 32.64$, $P = 0.03$; Training sessions: $F_{1, 2} = 0.54$, $P = 0.54$; Distribution x Training: $F_{1, 2} = 0.45$, $P = 0.57$. Overtrain action-initiation A2A⁺ neurons tend to respond on less than half the action-initiation events across training and this is consistent across training. **(f)** Within-session correlation of the activity around action initiation of each overtrain action-initiation excited neuron. 2-way ANCOVA, Neuron activity: $F_{1, 22.00} = 0.40$, $P = 0.53$; Training: $F_{1.24, 27.30} = 0.63$, $P = 0.47$; Activity x Time: $F_{1.23, 27.09} = 0.65$, $P = 0.46$. The activity of overtrain action-initiation excited A2A⁺ neurons around action initiation is not correlated above shuffled control within a training session. A2A-cre: $N = 4$ (1 male). Data presented as mean ± s.e.m. Males = closed circles, Females = open circles.



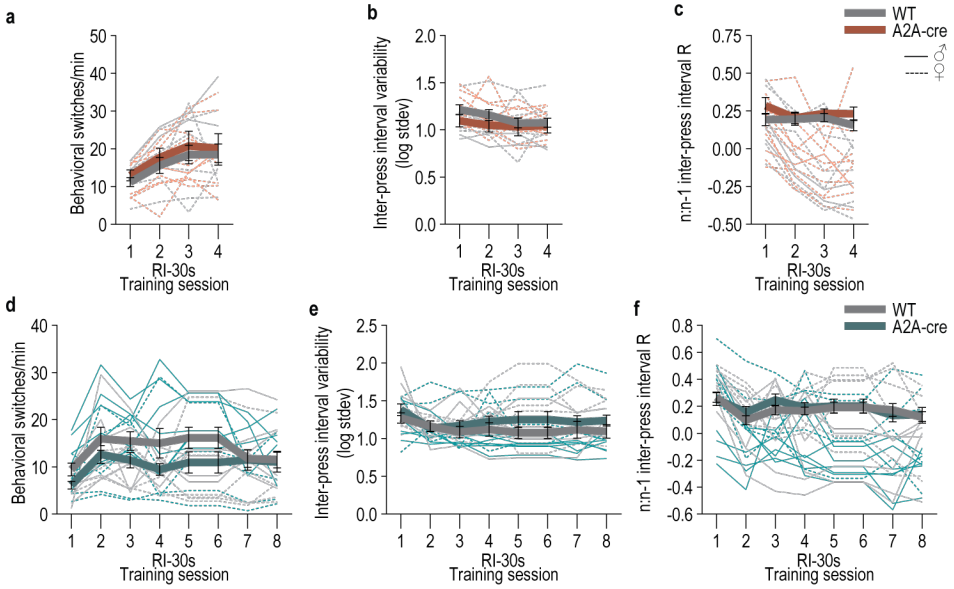
Extended Data Figure 5-7. Encoding of other task variables by A2A$^+$ neurons that drop out of the early action-initiation ensemble or get incorporated into the overtrain action-initiation excited ensemble. (a-b) Percentage of the neurons that dropout of the early action-initiation excited ensemble that are excited (a; 2-way ANOVA, Event x Training: $F_{1.79, 5.37} = 7.05$, $P = 0.03$; Event: $F_{1.21, 3.63} = 30.74$, $P = 0.006$; Training: $F_{1.13, 3.40} = 14.46$, $P = 0.02$) or inhibited (b; 2-way ANOVA, Event: $F_{1.98, 5.94} = 1.64$, $P = 0.27$; Training: $F_{1.50, 4.50} = 2.77$, $P = 0.17$; Event x Training: $F_{1.46, 4.37} = 1.25$, $P = 0.35$) by other events (terminating lever press, reward, food-delivery port check). By definition, these neurons stop being activated by action initiation with training. They do not begin to be activated by other task events with training. Instead, fewer of these neurons are activated by other task events and some of them become inhibited by tasks events. **(c-d)** Percentage of the neurons that are incorporated into the overtrain action-initiation excited ensemble that are excited (c; 2-way ANOVA, Training: $F_{1.1, 3.48} = 11.06$, $P = 0.03$; Event: $F_{2.24, 6.71} = 3.67$, $P = 0.08$; Event x Training: $F_{1.54, 4.62} = 2.41$, $P = 0.19$) or inhibited (d; 2-way ANOVA, Event: $F_{1.16, 3.48} = 1.95$, $P = 0.25$; Training: $F_{1.09, 3.28} = 1.24$, $P = 0.35$; Event x Training: $F_{1.89, 5.68} = 0.83$, $P = 0.48$) by other events (terminating lever press, reward, food-delivery port check). By definition, these neurons became activated by action initiation at overtraining. Earlier in training very few of these neurons were activated by other events. A small proportion of these neurons were inhibited by task events earlier in training and this proportion decreased to near 0 with training. Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines. *$P < 0.05$, **$P < 0.01$.


a A2A-cre DMS: AAV9-syn-Flex-jGCaMP7s
Lens
A2A+-GCaMP
b
c
dF/F
Raw GCaMP
Deconvolved
Action Initation
Action Termination
Reward
5 s
d Overtraining, 8 d RI-30s
Imaging A2A+ neurons
Press → Pellet
Test
Valued state
Press → Ø
Devalued state
Press → Ø
e
Early
Middle
End
Presses/min
FR1
RI-30s
Training session
f
Presses/min
Valued
Devalued
g
Devaluation index
h
Coregistered
Non-coregistered
Early, N = 221
Middle, N = 234
Overtrain, N = 202
d
Behavior class decoding accuracy (F1 score)
Initiating
Terminating
Reward
Port Check
e
Lever-press rate decoding accuracy (R value)
Early training
Middle training
Overtrain
Shuffled control
i
Early Action initation
Middle Action initation
Overtrain Action initation
Max
Min
Cell number
Time from action initation (s)
Early action-initation excited ensemble
Overtrain action-initation excited ensemble
n
Cell number
o
Time (s) from action initation
0.2 Z
p
Action initation
auROC
Time (s)
q
Average auROC
Time from action initation (0.5-s bin)
v
w
x
y
j
% of all neurons tracked
k
% of tracked neurons modulated by action initiation
r
Early:Middle
Middle:Overtrain
Cross-session neuron activity R
s
Cross-session population R
z
aa
l
% of tracked neurons excited by action initiation
m
% of tracked neurons inhibited by action initiation
t
Cross-session lever-press rate decoding accuracy (R value)
u
% of ensemble preserved
ab
ac

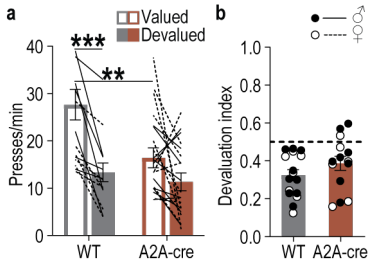
Extended Data Figure 5-8: The ensemble of DMS A2A$^+$ neurons that encodes action initiation is more stable if subjects remain goal-directed with overtraining. (a) Representative image of cre-dependent jGCaMP7s expression in DMS A2A$^+$ neurons. **(b)** Map of DMS GRIN lens placements for all subjects. **(c)** Representative dF/F and deconvolved calcium signal. **(d)** Procedure. RI, random-interval reinforcement schedule. **(e)** Training press rate. 2-way ANOVA, Training: $F_{1.49, 4.48} = 5.58$, $P = 0.06$. **(f)** Test press rate. 2-tailed t-test, $t_3 = 2.30$, $P = 0.11$, 95% CI -31.60 - 5.10. **(g)** Devaluation index [(Devalued condition presses)/(Valued condition presses + Devalued presses)]. One-tailed Bayes factor, $BF_{10} = 5.40$. **(h)** Representative recorded A2A$^+$ neuron spatial footprints during the first (early), 4th (middle), and 8th (overtrain) random-interval training session. Colored, co-registered neurons; white, non-co-registered neurons. **(i)** Heat map of minimum to maximum deconvolved activity (sorted by total activity) of each coregistered DMS A2A$^+$ neuron around lever-press action initiation. **(c)** Percent of all A2A$^+$ neurons coregistered across training in subjects that did not form habits with overtraining. 1-way ANOVA, $F_{1.32, 3.95} = 2.82$, $P = 0.17$. **(j)** Behavior class (initiating lever press, terminating press, reward collection, non-reinforced food-port check) decoding accuracy from A2A$^+$ coregistered neuron activity compared to shuffled control. Line at 0.25 = chance. 3-way ANOVA, Neuron activity (v. shuffled): $F_{1, 3} = 548.47$, $P < 0.001$; Training session: $F_{2, 6} = 4.11$, $P = 0.08$; Behavior class: $F_{3, 9} = 40.4$, $P < 0.001$; Neuron activity x Training: $F_{2, 6} = 5.95$, $P = 0.038$; Neuron activity x Behavior: $F_{3, 9} = 16.24$, $P < 0.001$; Training x Behavior: $F_{6, 18} = 2.55$, $P = 0.02$; Neuron activity x Training x Behavior: $F_{6, 18} = 5.11$, $P = 0.003$. **(k)** Lever-press rate decoding accuracy from A2A$^+$ coregistered neuron activity. R = correlation coefficient between actual and decoded press rate. 2-way ANOVA, Neuron activity: $F_{1, 3} = 195.5$, $P = 0.0008$; Training: $F_{1.64, 4.92} = 0.23$, $P = 0.77$; Neuron activity x Training: $F_{1.62, 4.86} = 0.23$, $P = 0.76$. **(l-m)** Percent of coregistered neurons (Average 106.5 coregistered neurons/mouse, s.e.m. 40.34) significantly modulated (f; 1-way ANOVA, $F_{1.05, 3.16} = 8.14$, $P = 0.06$), excited (g; 1-way ANOVA, $F_{1.14, 3.43} = 1.72$, $P = 0.28$), or inhibited (h; 1-way ANOVA, $F_{1.90, 5.70} = 3.37$, $P = 0.11$) around action initiation. **(n-q)** Activity and modulation across training of DMS A2A$^+$ early action-initiation excited neurons (N = 47 neurons/4 mice; average 9.4 neurons/mouse, s.e.m. = 12.44). Heat map (n), Z-scored activity (o), and area under the receiver operating characteristic curve (auROC) modulation index (p) of early action-initiation excited neurons around action initiation across training. (q) Modulation index averaged across 0.5-s bins around action initiation. 2-way ANCOVA, Training: $F_{1.82, 64.00} = 0.48$, $P = 0.62$; Time bin: $F_{2.18, 76.27} = 0.07$, $P = 0.94$; Training x Time: $F_{3.78, 132.35} = 1.61$, $P = 0.18$. **(r-s)** Cross-session correlation of the activity around action initiation of each early action-initiation excited neuron (r; 2-way ANCOVA, Neuron activity: $F_{1, 35} = 0.02$, $P = 0.90$; Training: $F_{1, 35} = 0.001$, $P = 0.98$; Activity x Time: $F_{1, 35} = 0.01$, $P = 0.91$) or the population activity of these neurons (s; 2-way ANOVA, Neuron activity: $F_{1, 2} = 31.38$, $P = 0.03$; Training: $F_{1, 2} = 1.25$, $P = 0.38$; Activity x Time: $F_{1, 2} = 1.21$, $P = 0.39$). **(t)** Cross-session decoding accuracy of lever-press rate from the activity of A2A$^+$ early action-initiation-excited neuron population activity on the 1st training session. Planned, Bonferroni corrected, 2-tailed t-tests, Early: $t_4 = 4.22$, $P = 0.04$, 95% CI 0.03 - 0.85; Middle: $t_4 = 1.97$, $P = 0.36$, 95% CI -0.21 - 0.61; Overtrain: $t_4 = 2.55$, $P = 0.19$, 95% CI -0.15 - 0.67. **(u)** Percent of A2A$^+$ early action-initiation excited neurons that continued to be significantly excited by action initiation on the 4th and 8th training sessions. 2-tailed t-test, $t_2 = 1.28$, P =0.33, 95% CI -85.86 - 46.47. **(v-y)** Activity and modulation across training of A2A$^+$ overtrain action-initiation excited neurons (N = 49 neurons/4 mice; average 9.80 neurons/mouse, s.e.m. 11.69). Heat map (v), Z-scored activity (w), and auROC modulation index (x) of overtrain action-initiation excited neurons around action initiation across training. (y) Modulation index around action initiation. 2-way ANCOVA, Training: $F_{1.65, 51.23} = 0.43$, $P = 0.43$; Time bin: $F_{2.16, 67.03} = 00.17$, $P = 0.86$; Training x Time: $F_{4.31, 133.64} = 0.59$, $P = 0.68$. **(z-aa)** Cross-session correlation of the activity around action initiation of each overtrain action-initiation excited neuron (a; 2-way ANCOVA, Neuron activity: $F_{1, 31} = 1.77$, $P = 0.19$; Training: $F_{1, 31} = 0.28$, $P = 0.60$; Activity x Time: $F_{1, 31} = 0.20$, $P = 0.66$) or the population activity of these neurons (aa; 2-way ANOVA, Neuron activity: $F_{1, 3} = 2.37$, $P = 0.22$; Training: $F_{1, 3} = 1.20$, $P = 0.35$; Activity x Time: $F_{1, 3} = 1.13$, $P = 0.37$). **(ab)** Cross-session decoding accuracy of lever-press rate from the activity of A2A$^+$ overtrain action-initiation-excited neuron population activity on the 8th training session. Planned, Bonferroni corrected, 2-tailed t-tests, Early: $t_6 = 1.48$, $P = 0.57$, 95% CI -0.16 to 0.41; Middle: $t_6 = 0.92$, $P > 0.99$, 95% CI -0.21 - 0.36; Overtrain: $t_6 = 4.73$, $P = 0.01$, 95% CI 0.12 - 0.69. **(ac)** Percent of A2A$^+$ overtrain action-initiation excited neurons that were also significantly excited by action initiation on 1st and 4th training sessions. 2-tailed t-test, $t_3 = 0.82$, P =0.47, 95% CI -32.59 – 55.21. A2A-cre: N = 4 (3 male). Data presented as mean ± s.e.m. Males = closed circles/solid lines, Females = open circles/dashed lines. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.



Extended Data Figure 6-1. Food-port entries during training and test with DMS A2A⁺ neuron chemogenetic manipulation. (a-c) Chemogenetic inactivation of DMS A2A⁺ neurons during learning. WT: N = 10 (1 male); A2A-cre: N = 7 (2 males). **(a)** Training entry rate. 2-way ANOVA, Training: $F_{2.87, 43.12} = 3.50$, $P = 0.02$; Genotype: $F_{1, 15} = 6.63$, $P = 0.02$; Training x Genotype: $F_{4, 60} = 0.31$, $P = 0.87$. **(b)** Test entry rate. 2-way ANOVA, Value x Genotype: $F_{1, 15} = 1.08$, $P = 0.32$; Value: $F_{1, 15} = 4.06$, $P = 0.06$; Genotype: $F_{1, 15} = 0.03$, $P = 0.86$. **(c)** Training average lever presses/earned reward outcome. 2-way ANOVA, Training: $F_{2.38, 35.73} = 32.06$, $P < 0.0001$; Genotype: $F_{1, 15} = 0.08$, $P = 0.78$; Training x Genotype: $F_{4, 60} = 0.47$, $P = 0.76$. **(d-f)** Chemogenetic activation of DMS A2A⁺ neurons during learning. WT: N = 11 (7 males); A2A-cre: N = 10 (6 males). **(d)** Training entry rate. 2-way ANOVA, Training: $F_{4.22, 92.89} = 6.10$, $P = 0.0002$; Genotype: $F_{1, 22} = 0.00008$, $P = 0.98$; Training x Genotype: $F_{8, 176} = 0.54$, $P = 0.83$. **(e)** Test entry rate. 2-way ANOVA, Value x Genotype: $F_{1, 22} = 1.50$, $P = 0.23$; Value: $F_{1, 22} = 0.35$, $P = 0.56$; Genotype: $F_{1, 22} = 0.27$, $P = 0.61$. **(f)** Training average lever presses/earned reward outcome. 2-way ANOVA, Training: $F_{2.42, 5334} = 16.33$, $P < 0.0001$; Genotype: $F_{1, 22} = 0.35$, $P = 0.56$; Training x Genotype: $F_{8, 176} = 0.76$, $P = 0.64$. **(g-i)** Chemogenetic inactivation of DMS A2A⁺ neurons during test of behavioral strategy after learning. WT: N = 16 (9 males); A2A-cre: N = 14 (8 males). **(g)** Training entry rate. 2-way ANOVA, Training: $F_{1.93, 53.92} = 28.28$, $P < 0.0001$; Genotype: $F_{1, 28} = 0.23$, $P = 0.63$; Training x Genotype: $F_{4, 112} = 0.70$, $P = 0.59$. **(h)** Test entry rate. 2-way ANOVA, Value x Genotype: $F_{1, 28} = 0.47$, $P = 0.50$; Value: $F_{1, 28} = 1.82$, $P = 0.19$; Genotype: $F_{1, 28} = 4.08$, $P = 0.05$. **(i)** Training average lever presses/earned reward outcome. 2-way ANOVA, Training: $F_{2.19, 61.42} = 81.93$, $P < 0.0001$; Genotype: $F_{1, 28} = 1.55$, $P = 0.22$; Training x Genotype: $F_{4, 112} = 2.25$, $P = 0.07$. Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines. Neither chemogenetic inhibition nor activation of DMS A2A⁺ neurons affected checks of the food-delivery port or altered the press-reward action-outcome relationship.



Extended Data Figure 6-2. Behavioral strategy variables during training with DMS A2A⁺ neuron chemogenetic manipulation. (a-b) Chemogenetic inactivation of DMS A2A⁺ neurons during learning. WT: *N* = 10 (1 male); A2A-cre: *N* = 7 (2 males). Behavioral switch rate (a; how often subjects stop one behavior to transition to another i.e., press → food-port check or food-port check → press; 2-way ANOVA, Training: $F_{2.30, 36.79} = 16.17$, $P < 0.0001$; Genotype: $F_{1, 16} = 0.02$, $P = 0.89$; Training x Genotype: $F_{2.30, 36.79} = 0.11$, $P = 0.92$), interpress interval variability (b; standard deviation of the log-transformed interpress-intervals; 2-way ANOVA, Training: $F_{2.64, 42.21} = 2.47$, $P = 0.08$; Genotype: $F_{1, 16} = 0.96$, $P = 0.34$; Training x Genotype $F_{2.64, 42.21} = 0.59$, $P = 0.60$), and relationship to past trial behavior (c; correlation between the current interpress interval (n) and the interpress interval of the prior press (n-1); 2-way ANOVA, Genotype: $F_{1, 16} = 1.27$, $P = 0.28$; Training: $F_{2.54, 40.70} = 1.15$, $P = 0.33$; Training x Genotype: $F_{2.54, 40.70} = 1.06$, $P = 0.37$) across days of training. **(c-d)** Chemogenetic activation of DMS A2A⁺ neurons during learning. WT: *N* = 11 (7 males); A2A-cre: *N* = 10 (6 males). Behavioral switch rate (d; 2-way ANOVA, Training: $F_{3.38, 74.41} = 4.51$, $P = 0.004$; Genotype: $F_{1, 22} = 2.39$, $P = 0.13$; Training x Genotype: $F_{3.38, 74.41} = 1.09$, $P = 0.36$), interpress interval variability (e; Training: $F_{2.97, 65.26} = 2.40$, $P = 0.08$; Genotype: $F_{1, 22} = 1.43$, $P = 0.24$; Training x Genotype: $F_{2.97, 65.26} = 0.66$, $P = 0.58$), and relationship to past trial behavior (f; 2-way ANOVA, Genotype: $F_{1, 22} = 0.21$, $P = 0.65$; Training: $F_{3.14, 69.01} = 3.34$, $P = 0.02$; Training x Genotype: $F_{3.14, 69.01} = 0.55$, $P = 0.65$) across days of training. Neither inhibition nor excitation of DMS A2A⁺ neurons affected behavioral strategy variables. Data presented as mean ± s.e.m. Males = solid lines, Females = dashed lines.



Extended Data Figure 6-3. Press rate during devaluation test with DMS A2A$^+$ neuron inhibition. (a) Test press rate. 2-way ANOVA, Value x Genotype: $F_{1, 28} = 6.46$, $P = 0.02$; Value: $F_{1, 28} = 28.86$, $P < 0.001$; Genotype: $F_{1, 28} = 5.15$, $P = 0.03$. **(b)** Devaluation index. 2-tailed t-test, $t_{28} = 1.36$; $P = 0.18$, 95% CI -0.03 - 0.16. Data presented as mean ± s.e.m. Males = closed circles/solid lines, Females = open circles/dashed lines. **$P < 0.01$, ***$P < 0.001$.

SUPPLEMENTAL TABLES

Subject ID	Genotype	Sex	Goal-directed or habitual	Total neurons			Co-registered neurons	% neurons coregistered		
				Early	Middle	End		Early	Middle	End
5330	D1cre	Female	Habitual	216	231	187	111	51.39%	48.05%	59.36%
5350	D1cre	Female	Habitual	204	236	250	146	71.57%	61.86%	58.40%
5399	D1cre	Female	Habitual	316	331	373	235	74.37%	71.00%	63.00%
6579	D1cre	Male	Habitual	134	171	154	55	41.04%	32.16%	35.71%
5543	A2Acre	Female	Habitual	111	137	148	74	63.06%	51.09%	47.30%
5567	A2Acre	Female	Habitual	85	98	78	30	35.29%	30.61%	38.46%
1932	A2Acre	Female	Habitual	352	329	285	113	32.10%	34.35%	39.65%
6620	A2Acre	Male	Habitual	123	133	147	80	65.04%	60.15%	54.42%
5562	A2Acre	Male	Goal-directed	38	55	49	14	36.84%	25.45%	28.57%
1786	A2Acre	Male	Goal-directed	221	234	202	169	76.47%	72.22%	83.66%
6608	A2Acre	Male	Goal-directed	308	322	321	179	58.12%	55.59%	55.76%
6619	A2Acre	Female	Goal-directed	126	168	167	64	50.79%	38.10%	38.32%

Supplemental Table 1: Neurons recorded and tracked for each subject.

		Prefeed Consumption		
		Valued Outcome mean +/- s.e.m.	**Devalued Outcome mean +/- s.e.m.**	**Statistics**
Figure 1	**D1-cre**	1.26 ± 0.23	1.20 ± 0.26	$t_3 = 0.54$ $P = 0.63$
Figure 1-1c	**Limited**	1.86 ± 0.14	2.28 ± 0.17	Training x Value: $F_{1, 14} = 0.35$, $P = 0.56$; Training: $F_{1, 14} = 12.95$, $P = 0.003$; Value: $F_{1, 14} = 3.40$, $P = 0.09$
	Overtrain	1.58 ± 0.14	1.79 ± 0.12	
Figure 3e	**WT**	0.96 ± 0.96	0.95 ± 0.19	Virus x Value: $F_{1, 14} = 0.27$, *P=0.61*; Virus: $F_{1, 14} = 0.026$, P=0.87; Value: $F_{1, 14} = 0.19$, P=0.68
	D1-cre	0.89 ± 0.14	0.94 ± 0.19	
Figure 3k	**WT**	1.50 ± 0.27	1.30 ± 0.23	Virus x Value: $F_{1, 11} = 0.92$, P=0.36; Virus: $F_{1, 11} = 0.50$, P=0.49; Value: $F_{1, 11} = 0.024$, P=0.88
	D1-cre	1.60 ± 0.25	1.70 ± 0.18	
Figure 3q	**WT**	1.36 ± 0.12	1.36 ± 0.14	Virus x Value: $F_{1, 18} = 0.032$, P=0.86; Virus: $F_{1, 18} = 0.087$, P=0.77; Value: $F_{1, 18} = 0.013$, P=0.91
	D1-cre	1.40 ± 0.20	1.43 ± 0.18	
A2A habitual vs Goal-directed	**A2A-cre habitual**	1.38 ± 0.24	1.25 ± 0.24	Group x Value: $F_{1, 6} = 0.06$, $P = 0.81$; Group: $F_{1, 6} = 0.13$, $P = 0.73$; Value: $F_{1, 6} = 0.83$, $P = 0.40$
	A2A-cre goal-directed	1.53 ± 0.07	1.30 ± 0.34	
Figure 4	**A2A-cre habitual**	1.38 ± 0.24	1.25 ± 0.24	$t_3 = 2.19$ $P = 0.12$
Figure 5-7	**A2A-cre goal-directed**	1.53 ± 0.07	1.30 ± 0.34	$t_3 = 0.59$ $P = 0.60$
Figure 6e	**WT**	1.35 ± 0.19	1.46 ± 0.22	Virus x Value: $F_{1, 16} = 1.05$, P=0.32; Virus: $F_{1, 16} = 0.03$, $P = 0.87$; Value: $F_{1, 16} = 0.0006$, P=0.98
	A2A-cre	1.50 ± 0.22	1.40 ± 0.19	
Figure 6k	**WT**	1.40 ± 0.18	1.30 ± 1.39	Virus x Value: $F_{1, 22} = 0.31$, P=0.58; Virus: $F_{1, 22} = 0.03$, P = 0.87; Value: $F_{1, 22} = 0.10$, P = 0.76
	A2A-cre	1.36 ± 0.09	0.16 ± 0.13	
Figure 6q	**WT**	1.82 ± 0.15	1.91 ± 0.14	Virus x Value: $F_{1, 28} = 0.56$, P=0.46; Virus: $F_{1, 28} = 3.5$, P=0.07; Value: $F_{1, 28} = 0.0056$, P=0.94
	A2A-cre	1.65 ± 0.14	1.53 ± 0.12	

Supplemental Table 2: Sensory-specific satiety prefeed consumption. Values reflect average amount consumed in grams ± s.e.m.

		Post-choice Consumption		
		Valued Outcome mean +/- s.e.m.	Devalued Outcome mean +/- s.e.m.	Statistics
Figure 1	**D1-cre**	0.14 ± 0.06	0.00 ± 0.02	$t_7 = 2.58$ $P = 0.04$
Figure 1-1c	**Limited**	0.32 ± 0.06	0.03 ± 0.02	Training x Value: $F_{1, 30} = 0.004$, $P = 0.95$; Training: $F_{1, 30} = 0.41$, $P = 0.53$; Value: $F_{1, 30} = 38.49$, $P < 0.001$
	Overtrain	0.29 ± 0.06	0.00 ± 0.00	
Figure 3e	**WT**	0.18 ± 0.04	0.06 ± 0.04	Group x Value: $F_{1, 30} = 0.038$, P=0.85; Group: $F_{1, 30} = 0.26$, P=0.62; Value: $F_{1, 30} = 12.22$, P=0.0015
	D1-cre	0.16 ± 0.06	0.03 ± 0.02	
Figure 3k	**WT**	0.14 ± 0.05	0.07 ± 0.03	Group x Value: $F_{1, 24} = 0.98$, P=0.33; Group: $F_{1, 24} = 0.21$, P=0.65; Value: $F_{1, 24} = 12.8$, P=0.0015
	D1-cre	0.15 ± 0.04	0.02 ± 0.02	
Figure 3q	**WT**	0.26 ± 0.04	0.05 ± 0.01	Group x Value: $F_{1, 38} = 0.047$, P=0.83; Group: $F_{1, 38} = 0.20$, P=0.66; Value: $F_{1, 38} = 57.41$, P<0.0001
	D1-cre	0.28 ± 0.05	0.06 ± 0.03	
A2A habitual vs Goal-directed	**A2A-cre habitual**	0.24 ± 0.10	0.01 ± 0.02	Group x Value: $F_{1, 14} = 0.11$, $P = 0.75$; Group: $F_{1, 14} = 0.26$, $P = 0.62$; Value: $F_{1, 14} = 12.22$, $P < 0.01$
	A2A-cre goal-directed	0.30 ± 0.11	0.02 ± 0.01	
Figure 4	**A2A-cre habitual**	0.24 ± 0.10	0.01 ± 0.02	$t_7 = 2.23$ P = 0.06
Figure 5-7	**A2A-cre goal-directed**	0.30 ± 0.11	0.02 ± 0.01	$t_7 = 2.72$ P = 0.030
Figure 6e	**WT**	0.15 ± 0.04	0.00 ± 0.01	Group x Value: $F_{1, 34} = 0.35$, P=0.56; Group: $F_{1, 34} = 0.012$, P=0.91; Value: $F_{1, 34} = 26.2$, P<0.0001
	A2A-cre	0.13 ± 0.03	0.01 ± 0.02	
Figure 6k	**WT**	0.19 ± 0.04	0.02 ± 0.04	Group x Value: $F_{1, 46} = 1.9$, P=0.18; $F_{1, 46} = 0.07$, P=0.79; Value: $F_{1, 46} = 29.2$, P<0.0001
	A2A-cre	0.16 ± 0.03	0.03 ± 0.02	
Figure 6q	**WT**	0.29 ± 0.04	0.07 ± 0.02	Group x Value: $F_{1, 58} = 0.55$, P=0.46; Group: $F_{1, 58} = 0.039$, P=0.85; Value: $F_{1, 58} = 72.62$; P<0.0001
	A2A-cre	0.31 ± 0.04	0.04 ± 0.02	

Supplemental Table 3: Average post-probe-test choice consumption. Values reflect average amount consumed in grams ± s.e.m.

Category	Item	Vendor	Catalog #	Lot #	Titer
Viruses	AAV2-hSyn-DIO-hM3D(Gq)-mCherry	Addgene	44361	v97910	2.0×10^{13} gc/mL
	AAV2-hSyn-DIO-hM4D(Gi)-mCherry	Addgene	44362	v68359	1.5×10^{13} gc/mL
	AAV9-Syn-FLEX-GCaMP7s	Addgene	50459	v107704	1.6×10^{13} gc/mL

Supplemental Table 4: Key reagents information.