Striatal dopamine encodes the relationship between actions and reward

Authors: G. Hart[1*], T.J. Burton[1*] & B.W. Balleine[1#]

Affiliations:

1. Decision Neuroscience Lab, School of Psychology, UNSW Sydney

* These authors contributed equally to this work.

Corresponding Author: bernard.balleine@unsw.edu.au

Abstract:

Despite the well-described role of dopamine in the striatum for skill learning and movement, its function in encoding actions and their outcomes for goal-directed action is less clear. Here, rats acquired two actions for distinct outcomes while we simultaneously recorded dopamine release in the dorsomedial striatum as these action-outcome associations were incremented and selectively degraded. Goal-directed actions generated a lateralized dopamine signal that reflected the strength of the action-outcome association and tracked increments and decrements in the specific action-outcome relationship, which was sensitive to changes in both contingency and outcome identity. This lateralized signal was updated by a dopamine action value signal that was broadcast bilaterally during the action and updated by a reward prediction error following exposure to the outcome. Our results establish, therefore, a dual role for striatal dopamine signaling during goal-directed actions: bilateral trial-by-trial burst signaling encoding action values, which modifies the long-term action-outcome association, reflected in emergent lateralization of dopamine release during actions.

The capacity for goal-directed action is a core function that allows animals to encode the consequences or outcome of their actions and so make flexible choices to maintain adaptive behavior in a changing environment(Dickinson & Balleine, 1994; Dolan & Dayan, 2013). Recent evidence suggests that action-outcome encoding depends on a prefronto-striatal circuit focused on the posterior dorsomedial striatum (pDMS)(Balleine, 2019; Balleine & O'Doherty, 2010; Hart, Bradfield, & Balleine, 2018; Hart, Bradfield, Fok, et al., 2018) with initial learning and the subsequent updating of these associations involving changes in plasticity at two types of principal neuron(Balleine et al., 2021): the striato-nigral direct spiny projection neurons (dSPNs), which express dopamine D1 receptors, and striato-pallidal indirect SPNs (iSPNs) expressing the D2 receptor(Gerfen & Surmeier, 2011) (Matamales et al., 2020; Peak et al., 2020). Importantly, this plasticity appears to reflect the integration of glutamatergic inputs from cortical, thalamic and limbic regions with the input from midbrain dopaminergic neurons(Bradfield et al., 2013; Fisher et al., 2020; Holly et al., 2019; J. Lee et al., 2020; Nonomura et al., 2018; Peters et al., 2021; Reynolds et al., 2001; Shiflett & Balleine, 2011). Nevertheless, although much has been learned about the role of striatal dopamine in the cellular and circuit level plasticity necessary for reflexive movements and skills (Klaus et al., 2017; Park et al., 2020; Santos et al., 2015) relatively little is known about its role in the acquisition and performance of goal-directed actions. Here we show that over the course of acquisition, goal-directed actions develop a lateralized dopamine signal that reflects the strength of the action-outcome association, tracking increments and decrements in the action-outcome contingency. This signal was updated by a dopamine dependent action value signal, broadcast bilaterally during an action, and updated by a reward prediction error induced by exposure to the outcome. These findings demonstrate that bilateral dopamine release during goal-directed action conveys anticipated action value allowing the strength of the action-outcome association to be updated, the latter reflected in the magnitude of a lateralized dopamine signal. We establish, therefore, that dopamine plays a core modulatory role in the acquisition and updating of goal-directed action.

Dopamine release becomes lateralized over the course of goal-directed learning.

We utilized a genetically-encoded dopamine indicator, dLight1.1(Patriarchi et al., 2018) (**Fig 1A, Fig S1A**), to measure rapid changes in dopamine release in the pDMS during the acquisition and performance of goal-directed actions. This indicator inserts circularly permutated GFP into dopamine D1 receptors such that the binding of dopamine causes an increase in fluorescence, which we measured using fiber photometry. We measured dopamine release during initial learning and subsequently during performance and then assessed effects specific to changes in the strength of the action-outcome association when this was updated during a contingency degradation treatment.

Rats were trained on a goal-directed instrumental conditioning protocol in which two levers, located to the left and right of a central food port or magazine, each delivered distinct rewards into the central magazine (grain pellets and sucrose solution) on increasing random interval (RI) schedules of reinforcement across sessions (**Fig 1B**). Response rates on each lever and mean entry rates across days of training are presented in **Fig 1C**. Levers were labelled according to subsequent treatment as to-be-degraded or to-be-non-degraded, with side of lever (left and right) counterbalanced within each condition. Rats increased responding on both levers across

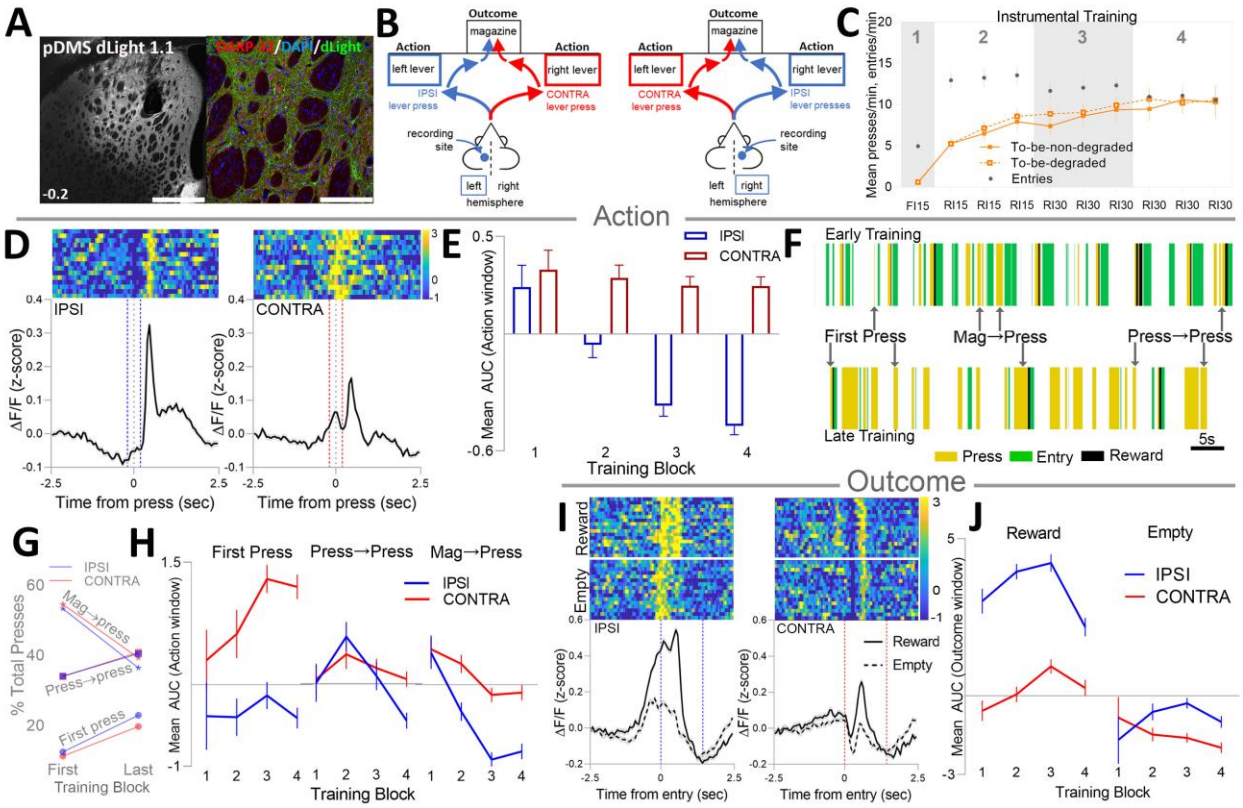

Fig. 1. Dopamine release in the pDMS is lateralized during goal-directed actions and their outcomes. (**A**) Left: Location of fiber cannula and expression of dLight1.1 in the pDMS (0.2mm posterior to bregma), scale bar =1000µm. Right: from the same slice, dLight1.1 (green), DARP-32 (red) and DAPI (blue), scale bar=200µm. (**B**) Rats were trained with alternating levers left and right of the magazine, designated as IPSI and CONTRA, relative to the recording hemisphere. (**C**) Mean (±SEM) presses per minute on each lever and entries per minute across each day of training, numbers in grey specify training blocks. (**D**) ΔF/F dLight signals over time aligned to IPSI (left) and CONTRA (right) lever presses (grey line). Top: Example trials showing color-coded ΔF/F signals. Bottom: the mean ΔF/F averaged across all actions during instrumental training, Action Window indicated by colored lines. (**E**) The mean (+ SEM) area under the curve (AUC) of the ΔF/F signal during the Action Window for all IPSI and CONTRA lever presses, averaged within each training block. (**F**) Behavioral sequence over 2 min for one rat during Block 1 of training (top) and the same rat in Block 4 of training (bottom), lever presses in yellow, entries in green and rewards in black, press types identified. (**G**) Each press type represented as a percent of total presses on the first and last day of training (**H**) The mean (± SEM) AUC of the ΔF/F signal for IPSI and CONTRA presses, subdivided according to press type, averaged within each training block. (**I**) ΔF/F dLight signals over time aligned to rewarded and non-rewarded (empty) magazine entries after IPSI (left) and CONTRA (right) presses. Top: Example trials showing color-coded ΔF/F signals. Bottom: the mean ΔF/F averaged across all rewarded and empty magazine entries after presses during instrumental training, Outcome Window indicated by colored lines. (**J**) Mean (± SEM) AUC of the ΔF/F signal during the Outcome Window for all rewarded (left) and empty (right) magazine entries after IPSI and CONTRA presses, averaged within each training block.

days (significant linear trend F(1,20)=62.5, p= 1.40e-7), and there were no differences in the response rates of each action (F's<1.0).

To identify epochs of interest for photometry analysis, we compiled the temporal dynamics of all lever press-magazine entry sequences (**Fig S1B**) and sought to analyze non-overlapping windows around the Action (lever press) and Outcome (i.e., when rats either obtained or failed to obtain a food reward). The Action window extended 0.2s before to 0.2s after a lever press, whereas the Outcome window captured entry to the reward magazine area and began on entry after a press and continued for 1.5s (see **Fig S1B**). We recorded dopamine release in the pDMS in one hemisphere (left or right, counterbalanced), with presses categorized according to whether the lever was located ipsilateral (IPSI) or contralateral (CONTRA) to the recording hemisphere (**Fig 1B)**.

The mean average z-scored $\Delta F/F$ signals aligned to IPSI and CONTRA presses averaged across instrumental training are presented in **Fig 1D**. There was a clear difference in the profile of dopamine release during IPSI relative to CONTRA presses. To quantify the relative change in release over acquisition, we subdivided the training into 4 blocks (indicated in **Fig 1C**) and calculated the area under the curve (AUC; trapezoidal method) within each Action window (**Fig 1E).** Dopamine release was progressively suppressed during IPSI presses but was maintained during CONTRA presses across training. There was no difference in release during IPSI and CONTRA presses in Block 1 (p=0.6), but a significant difference in Blocks 2-4 (k=6, p-critical=0.008, p's < 0.0002).

Next, we classified IPSI and CONTRA presses according to their location within press-entry sequences; **Fig 1F** shows a subset of behavioral sequences for one rat early and late in training. As training developed, presses and entries became loosely clustered into press-entry sequences interspersed with single presses. We therefore subdivided presses into three categories according to the immediately preceded event (based on a time window that was 1 standard deviation of the inter-press-interval on that lever for that session): (1) First Presses, for which neither a press nor an entry had occurred in the preceding window, also referred to as sequence initiations; (2) Press→Presses which followed another press in sequence; and (3) Mag→Presses, which followed a magazine entry in sequence. The total number of each press type in each training block is presented in **Fig S1C** which shows there was an increase in First Presses and Press→Presses across training and a decrease in Mag→Presses (**Fig 1G**). The mean AUCs for each action category are presented in **Fig 1H.** Dopamine release was significantly lateralized during First presses (k=15, p-critical=0.003, IPSI vs CONTRA p=1.61E-25) and Mag→Presses (IPSI vs CONTRA p=1.75E-17), but not Press→Presses (IPSI vs CONTRA p=0.02), however this difference developed across training; there were no differences between IPSI and CONTRA presses of any type in Block 1 (p's >0.2).

To assess whether this emergent lateralization was related to increased response vigor as training progressed, we assessed the correlation between the session-average inter-press interval against the session-average AUC during the Action window for each press type across each training block (**Fig S1D**). There were no significant correlations between press rate and dopamine release in any press type in any session (p's>0.3). Next, we assessed whether lateralization was related to the individual inter-event interval (i.e. the interval between the

press and its immediately preceding press or entry); the correlations between the inter-event-interval for each press and its action window AUC are presented in **Fig S1E**. There was no correlation between dopamine release and inter-event interval for IPSI or CONTRA First Presses or Press→Presses (p's>0.1) however there was a positive relationship between both IPSI and CONTRA Mag→Presses and the time since their preceding magazine entry in the final blocks of training (**Fig S1E-right panel**). For Mag→IPSI presses, this relationship indicated that shorter entry→press latencies were associated with greater dopamine inhibition, whereas, for Mag→CONTRA presses, shorter latencies were associated with less dopamine release.

We also assessed whether dopamine release in the Action window differed according to the event following its performance: **Fig S1C** (top) shows the percentage of each press type followed by another press (Press-Press), a non-rewarded magazine entry (Press-Empty) or a rewarded magazine entry (Press-Rew). Press sequences (Ipsi→IPSI or Contra→CONTRA) were least likely to be rewarded, and most likely to be followed by another press, whereas first presses (i.e. First IPSI or First CONTRA) were most likely to be rewarded. The mean AUCs during the Action window for each of these press subtypes are presented in **Fig S1F.** For each press subtype, there were no differences in dopamine release during actions followed by either another press or an empty magazine check (k=18, p-critical=0.003, p's > 0.15). There was, however, significantly greater release during rewarded actions that had followed magazine entries (Mag→IPSI and Mag→CONTRA, p's < 0.0002; p's > 0.01 for other press subtypes), suggesting that reward predictions were enhanced at such moments.

Dopamine release during outcome retrieval reflects both movement and reward

To establish whether this lateralized dopamine signal was specific to the Action window, we assessed dopamine release during the window 0.2-0.4s after the press, during which the reward was delivered (if there was reward) and animals often turned in the opposite direction from the lever to approach the magazine (illustrated in **Fig 1B**). These responses were categorized according to both the preceding press and the immediately following outcome: i.e., whether IPSI and CONTRA presses followed by magazine entries were rewarded or not (Press→Rew, Press→Empty), or were followed by other presses with no magazine entry (Press→no entry). The mean AUCs across this Magazine Approach window are presented in **Fig S1G**. There was a pronounced dopamine release spike in this window on rewarded trials, corresponding to the reward delivery (starting 0.2s after press), the magnitude of which increased across training blocks (k=52, p-critical=0.001; IPSI→Rew and CONTRA→Rew, Block 1 vs Block 4 p's < 1.04E-06; other press types Block 1 vs Block 4 p's > 0.06). However, importantly, whereas there was heightened dopamine release during CONTRA presses (relative to IPSI presses) during the Action window, this pattern reversed during the Magazine Approach window: dopamine release was significantly higher following IPSI presses, as rats made the contralateral turn towards the magazine (IPSI→Rew vs. CONTRA→Rew, p's < 0.0008 across all training blocks). Therefore, dopamine release increased across training as rats learned to detect the reward delivery and to approach the magazine, and this was augmented when the approach involved a contralateral turn from the lever to magazine, in accord with the general lateralization of release during movement. Importantly, therefore, the lateralization of dopamine during the lever press was

temporally restricted to the Action window and clearly distinguishable from subsequent dopamine signals associated with magazine approach.

Finally, we assessed dopamine release during the Outcome window. The mean z-scored $\Delta F/F$ signals averaged across instrumental training aligned to magazine entry are presented in **Fig 1I**; the dopamine release profile differed for entries that followed IPSI vs. CONTRA presses. The relative dopamine release averaged across the Outcome window (0-1.5s after entry onset) is reflected in the mean AUCs (**Fig 1J**), showing significantly greater dopamine release during rewarded than empty magazine entries (k=32 p-critical =0.0015, IPSI→Rew vs. IPSI→Empty Blocks 1-4, p's<5.8E-05; CONTRA-Rew vs CONTRA-Empty Blocks 2-4 p's<0.0002). The direction of the immediately preceding turn also significantly modulated the magnitude of dopamine release during magazine entry: the contralateral turn after an ipsilateral press resulted in greater dopamine release than the ipsilateral turn after a contralateral press (IPSI→Rew vs. CONTRA→Rew Blocks 1-4 p's <8.6E-07, IPSI→Empty vs. CONTRA→Empty Blocks 3-4 p's<0.0012, note there are very few empty magazine entries in Block 1).

Dopamine release during instrumental acquisition reports the updated action value

Our results suggest that during instrumental training, dopamine release is consistently greater in the hemisphere contralateral to movement, however the evolution of this signal across training differs in each stage of the action-outcome sequence: Dopamine release diverges between the ipsilateral and contralateral hemispheres during the action, whereas, during outcome retrieval, it is lateralised from the outset of training and remains relatively unchanged as training progresses. As such, these results suggest that changes in dopamine release across training reflect learning engaged by the action rather than the outcome. To assess this hypothesis more directly, we next sought to examine whether dopamine release during goal-directed actions is modulated by a reward prediction error generated within the Action and Outcome windows described above.

We calculated reward prediction-errors for each action-outcome association across training for each animal, whereby action values (V) were updated using a simplified reinforcement learning rule(Sutton & Barto, 1998). According to this model, at the time of magazine entry, prediction error is calculated, which is used to update the subsequent action value according to: $V(t+1)=V(t)+\delta(t)$ (**Fig 2A**), where δ is the prediction error modulated by a learning rate parameter (α) according to: $\delta(t)=\alpha[R(t)-V(t)]$, where R is the reward value, which is set at 0 for no reward or 1 for reward to generate a negative δ on non-rewarded entries and a positive δ on rewarded entries. Importantly, this simplified model assumes that reward predictions are only updated by experience with the consequences of the action (i.e., a rewarded or empty magazine entry); therefore, during press sequences in which the rat doesn't enter the magazine V will remain unchanged.

Fig 2B shows the mean z-scored $\Delta F/F$ signals during the Action window across a range of V values (low-high: 0-1) corresponding to the calculated action value (V) at the time of the action. The mean AUCs during the Action window for each action subtype are presented in **Fig 2C.** There was a clear relationship between dopamine release and V but only for Mag→Presses (k=6, p-critical=0.008; V(low) vs. V(high): Mag→IPSI, p=2.21E-22, Mag→CONTRA p=2.79E-12),

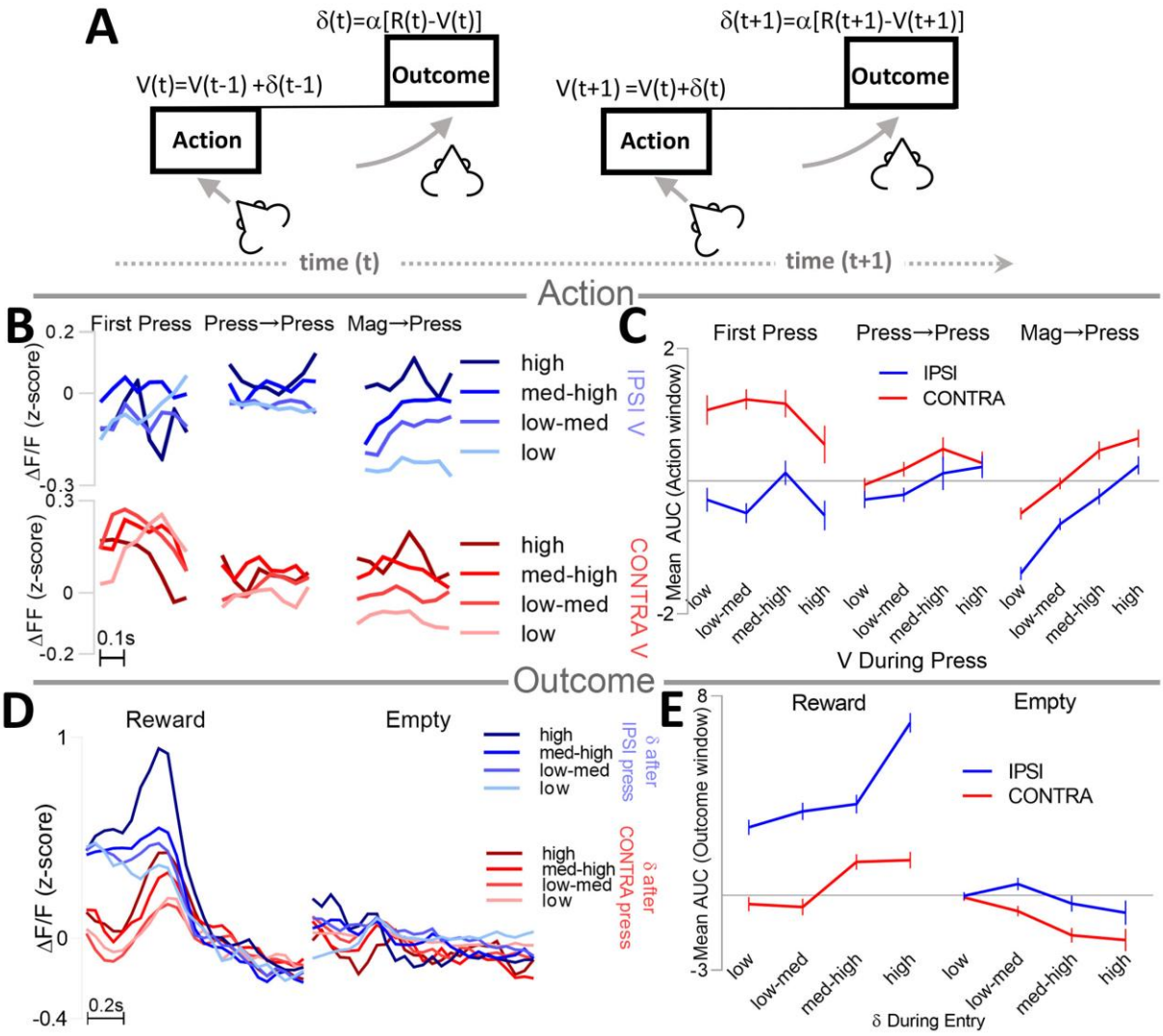

Fig. 2. Dopamine reflects action values and outcome reward prediction errors. (**A**) Action values are updated by summing the previous action value with the reward prediction error calculated during the preceding magazine entry (Outcome). (**B**) Mean ΔF/F signals during the Action window (-0.2s before press to 0.2s after press) for IPSI (blues) and CONTRA (reds) press subtypes, subdivided according to action value. (**C**) Mean AUC (±SEM) for signals shown in (B). (**D**) Mean ΔF/F signals during the Outcome window (0s to 1.5s after entry) for rewarded (left) and empty (right) magazine entries after IPSI (blues) and CONTRA (reds) press subtypes, subdivided according to reward prediction error calculated during that same entry. (**E**) Mean AUC (±SEM) for signals shown in (D).

and not for First Presses or Press→Presses (V(low) vs. V(high): First Presses p's>0.1; Press-Presses p's > 0.02). Note that, as our model assumes that prediction error is only generated during magazine entries, this finding suggests that these changes in dopamine release reflect the updating of action values. However, to ensure that this assumption within the model didn't drive the pattern of results, we modified the model to allow V to be updated on every press. In this case, all non-rewarded presses resulted in negative δ (i.e., presses followed by empty magazine checks and presses followed by other presses), and only presses followed by rewarded entries accrued a positive δ. The mean AUCs during presses across binned values of V using this model are presented in **Fig S2A**. Although there was generally more variability, the findings remained the same: dopamine release during presses that had followed magazine entries was positively related to V (k=6, p-critical=0.008; V(low) vs. V(high): Mag→IPSI p=3.36E-08, Mag→CONTRA p=0.0008), and there was no significant relationship between dopamine release and V during other press subtypes (i.e. V(low) vs. V(high): First Presses and Press→Presses, p's > 0.05).

Dopamine release elicited by the outcome reports a reward prediction error

The previous results confirm that dopamine release during the action in the pDMS most consistently reflects the value of the goal-directed action but only for the action following a magazine entry, suggesting that it was only after contact with reward (or non-reward) – and the consequent exposure to the prediction error – that action values are updated. To further examine this finding, we analyzed the magnitude of both reward prediction error (δ) and dopamine release during the Outcome window. These data are presented in **Fig 2D-E**. We binned magazine entries according to the magnitude of the error (δ) generated by the model (high-low), assuming δ is only calculated during entries. The mean z-scored ΔF/F signals during the Outcome window across differing δ values, which are positive for rewarded entries and negative for unrewarded entries, are presented in **Fig 2D.** As previously noted, the release profile during the Outcome window differs for entries following IPSI and CONTRA presses and was maintained across varying δ values. During the release peak for rewarded entries (**Fig 2D, left**), there was a graded increase in dopamine release according to δ and this was reflected in the mean AUCs, **Fig 2E.** There was a positive relationship between dopamine release and positive prediction errors (δ) during the Outcome window when rewards were delivered (k=4 p-critical=0.01, δ(low) vs. δ(high): IPSI→Rew p=2.35E-17; CONTRA→Rew p=5.22E-05). Equally importantly, there was a significant relationship between negative prediction errors and unrewarded magazine entries (δ(low) vs. δ(high): CONTRA→Empty p=0.0005), however this relationship wasn't significant for IPSI entries (IPSI→Empty, p=0.16). To directly compare waveform profiles for retrieved outcomes with high δ vs. low δ, we used a bootstrapping confidence intervals procedure (Methods, **Fig S2B**), which confirmed that dopamine release was significantly greater for (positive) high δ vs. low δ during the release peak for rewarded entries and significantly lower for (negative) high δ vs. low δ during the comparable timepoint for nonrewarded CONTRA entries. By contrast, for IPSI entries, dopamine release was inversely related to δ (negative) at the very start of the Outcome window (i.e., δ(high) > δ(low)) and this pattern reversed to reflect the negative prediction error signal at the end of the window (i.e., δ(low) > δ(high)). Together these results confirm that dopamine release during reward reports a positive prediction error and during non-reward a negative prediction error.

Contingency degradation reverses Action-specific but not Outcome-specific hemispheric lateralization.

If dopamine lateralization reflects the action-outcome association then it should increase during acquisition and decline when the action-outcome contingency is degraded. To investigate the relationship between dopamine lateralization and the strength of the action-outcome association more directly, we measured release during the degradation of one of the two action-outcome contingencies. Both levers continued to be presented in the same fashion and earned the same outcomes on the same interval schedule as during training, however, in addition, throughout each session (both levers) rats received non-contingent, i.e., response independent, deliveries of one of the earned outcomes delivered on a random time (RT) schedule (**Fig 3A**). It has previously been demonstrated that, for goal-directed actions, response independent outcomes 'degrade' previously established associations between an action and that specific outcome, resulting in the selective reduction in performance of the action relative to actions delivering other outcomes(Colwill & Rescorla, 1986; Dickinson & Charnock, 1985; Rescorla, 1992). Responding relative to baseline (Block 4 of instrumental training, see Methods) on the degraded and non-degraded levers across the 10 days of contingency degradation is presented in **Fig 3B**. We compared responding on the degraded and non-degraded levers averaged across training blocks (i.e. Blocks 1-5, 2-5, 3-5, 4-5 and 5 only). There were no differences between response rates on the two levers averaged across all days of training (Deg vs. nonDeg Blocks 1-5 $p>0.9$), however responding on the degraded lever was gradually suppressed from Block 3 onward, confirmed by a marginal difference between response rates on the degraded and non-degraded levers on training Blocks 4-5 ($k=5$, p-critical=0.01, $p=0.02$), and a significant difference between degraded and non-degraded levers in Block 5 ($p=0.005$). **Fig 3C** shows subset of behavioral sequences for one rat on the degraded lever early and late in degradation training, and **Fig 3D** shows the proportion of each press type represented in the total presses made. Despite an overall reduction in press rate on the degraded lever, the relative proportions of each press type were similar for the degraded and nondegraded lever; there was an increase in Mag→Presses, a decrease in Press→Presses and no change in First Presses for both levers.

We measured dopamine release in the pDMS across the 10 days of contingency degradation training. The mean AUCs over the Action window for IPSI and CONTRA presses that were degraded or non-degraded are presented across training blocks in **Fig 3E** which shows that the effect of degradation on dopamine release was apparent on both IPSI and CONTRA presses: Degradation increased dopamine release during IPSI presses, and reduced dopamine release during CONTRA presses, and this effect emerged across degradation training. Specifically, there was no difference between degraded and nondegraded levers during the first training block **(Fig 3B**; IPSI and CONTRA Deg vs. nonDeg, Block 1 p's > 0.1), but a significant difference across all subsequent blocks for IPSI presses ($k=10$, p-critical = 0.005, IPSI Deg vs. nonDeg, Blocks 2-5, p's <0.0001). There was also a significant effect of contingency degradation for CONTRA presses in Blocks 2 and 4 (CONTRA Deg vs. nonDeg, p's <3.18E-05), but not for Blocks 3 and 5 (p's >0,1). To assess the possible source of this variability, we analyzed each press type separately across the last 4 blocks of training, **Fig 3F-G**. The effect of contingency degradation was to reverse the hemispheric lateralization observed for First Presses, while maintaining hemispheric

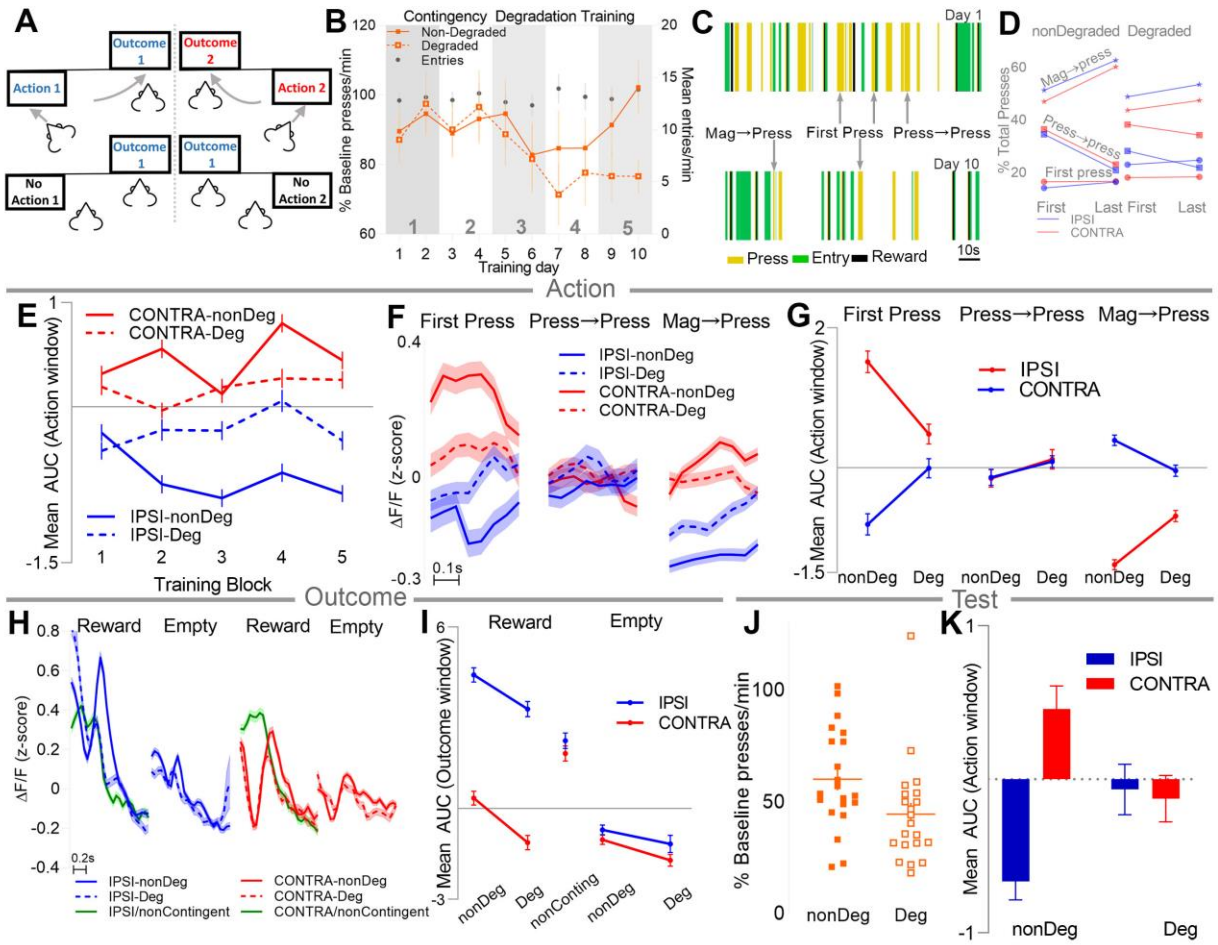

Fig. 3. Dopamine lateralization during the Action but not the Outcome reflects instrumental contingency. (**A**) In addition to each action-outcome pairing, rats received noncontingent (i.e. response-independent) deliveries of one of the earned outcomes. (**B**) Mean (±SEM) percent of baseline presses per minute on the non-degraded (nonDeg) and degraded (Deg) levers and mean (±SEM) entries per minute (Entries) across 10 days of contingency degradation training. Numbers in grey indicate training blocks. (**C**) Behavioral sequence over 2 min for one rat during Day 1 of degradation training (top) and the same rat on Day 10 (bottom), lever presses in yellow, entries in green and rewards in black, press types identified. (**D**) Each press type represented as a percent of total presses on the first and last day of degradation training. (**E**) Mean (±SEM) AUC of the ΔF/F signal during the Action window for IPSI (blue) and CONTRA (red) presses that were degraded or non-degraded, averaged across training blocks of contingency degradation. (**F**) ΔF/F signals during the Action window for IPSI (blue) and CONTRA (red) press subtypes, averaged across Blocks 2-5 of contingency degradation training. Thin lines indicate ±SEM. (**G**) Mean (±SEM) AUC for signals shown in (F). (**H**) ΔF/F signals during the Outcome window for rewarded and empty entries following IPSI (blue) and CONTRA (red) presses, and during retrieval of noncontingent rewards (green), averaged across degradation training. (**I**) Mean (±SEM) AUC of the ΔF/F signal during the Outcome window, averaged across contingency degradation training. Left: Rewarded entries that were contingent on IPSI (blue) or CONTRA (red) presses that were degraded (Deg) or nondegraded (nonDeg), or noncontingent rewards (nonConting). Right: Empty entries following degraded or nondegraded IPSI or CONTRA presses. (**J**) Percent of baseline presses per minute on the degraded and non-degraded levers for each animal during the contingency degradation test. Lines indicate mean ±SEM. (**K**) Mean (±SEM) AUC of the ΔF/F signal during the Action window for IPSI (blue) and CONTRA (red) presses that were degraded or nondegraded, averaged across the contingency degradation test.

lateralization during non-degraded actions (k=6 p-critical=0.008; First Press IPSI vs CONTRA; nonDeg p=1.19E-25, Deg p>0.01). Although lateralization was diminished for degraded Mag→Presses, a significant difference between IPSI and CONTRA actions remained (Mag→Presses IPSI vs CONTRA; nonDeg p=1.81E-64, Deg p=1.54E-08). Consistent with training, there was no evidence of hemispheric lateralization in either degraded or non-degraded press sequences (Ipsi→IPSI vs. Contra→CONTRA, p's>0.1).

Next, we assessed dopamine release during the Outcome window, presented in **Fig 3H-I**. The distinct profiles of dopamine release during magazine entries following IPSI and CONTRA presses (**Fig 3H)** were maintained for degraded presses, and critically, these shapes looked very different for entries following noncontingent outcome deliveries (IPSI/noncontingent and CONTRA/noncontingent), which (by design) didn't follow a press but were delivered outside of IPSI or CONTRA presses. The mean AUCs during the Outcome window following each press type and following noncontingent rewards are presented in **Fig 3I**. There was no difference in dopamine release during response-independent rewards delivered during IPSI or CONTRA lever blocks (IPSI/noncontingent vs. CONTRA/noncontingent, k=9, p-critical=0.006, p=0.13), however there was a general reduction in release during degraded outcome retrieval: For empty magazine checks this effect wasn't significant (Deg vs. nonDeg: IPSI→Empty p=0.1; CONTRA→Empty, p=0.0064), however, this reduction was significant for degraded rewards (IPSI→Rew and CONTRA→Rew, Deg vs. nonDeg, p's < 0.002), indicating that contingency degradation significantly depressed reward-related dopamine release.

To confirm that contingency degradation training induced a long-term change in the action-outcome association and not just within-session suppression, we conducted a choice extinction test on the day after the final contingency degradation training session (**Fig 3J**). Despite no rewards being delivered during this test, the suppression of performance on the degraded lever was maintained, with rats pressing the degraded lever significantly less than the non-degraded lever (k=1, p-critical=0.05, p=0.007), confirming the contingency degradation effect. We measured dopamine release during this extinction test, presented in **Fig 3K.** The pattern of results observed during the Action window across contingency degradation training was maintained on test: There was significantly greater dopamine release (AUC) during CONTRA presses relative to IPSI presses on the nondegraded lever (k=4, p-critical = 0.01, IPSI-nondeg vs CONTRA-nonDeg, p=3.95E-09), however this difference was abolished on the degraded lever (IPSI-Deg vs CONTRA-Deg p>0.7). As in training, the loss of lateralization during these actions was due to a bidirectional change: There was significantly less dopamine released during degraded than non-degraded CONTRA presses (CONTRA-nonDeg vs CONTRA-Deg p=0.007), and significantly more dopamine (or, more accurately, significantly less inhibition) during degraded IPSI presses (IPSI-nonDeg vs IPSI-Deg p=0.003). Therefore, the action-specific lateralization of dopamine release was significantly reversed when the instrumental relationship was degraded, and this change was preserved under extinction conditions when no outcomes were delivered.

Dopamine lateralization is sensitive to outcome identity

Having established that the dopamine lateralization is sensitive to the contingency between the action and reward, we next sought to establish whether this phenomenon was likewise impacted by changes in the outcome identity. Goal-directed action-outcome associations are

underpinned by a relationship between the action and its specific outcome goal. If dopamine lateralization reflects the specific action-outcome association underlying goal-directed action, it should be apparent across different reinforcement schedules (i.e., across both ratio and interval schedules) and it should be sensitive to changes in outcome identity. To test this, we injected dLight1.1 into the pDMS of a naïve cohort of rats (**Fig S3A**) and trained them on the same goal-directed instrumental training protocol described, except the rewards were delivered on a random ratio (RR) schedule of reinforcement (**Fig 4A**). As should be expected, this training schedule resulted in the development of tighter chunks of press sequences, which generally terminated with reward delivery and entry to the magazine (**Fig 4B**). In contrast to interval schedule training, in which Mag→Presses and Press→Presses contributed similarly to the total presses, ratio training resulted in the dominance of Press→Press sequences (**Fig 4C, Fig S3B**).

Response rates across instrumental training are presented in **Fig 4D** which increased across days (linear trend $F(1,5)=62.05$, $p=0.001$) and did not differ between IPSI and CONTRA levers ($F < 1.0$). Although ratio schedule training produced a different profile of lever press responding to interval schedules, the lateralized release of dopamine in the pDMS was similar; the mean AUC during the action window for each press type is shown across training blocks in **Fig 4E**. There was significant hemispheric lateralization (CONTRA > IPSI) during First Presses and Mag→Presses (k=6, p-critical=0.008, IPSI vs CONTRA averaged across Blocks 1-4; First Press p=2.55E-05, Mag→Press p=3.11E-20), but not for Press→Presses (p=0.3). Importantly, and consistent with what we observed previously, this lateralization developed across training; and there were no differences between IPSI and CONTRA presses of any type in Block 1 (p's>0.03).

We next assessed whether there was a correlation between (i) the session-based press rate and (ii) the latency to press after a press or entry, and dopamine release during IPSI and CONTRA presses, shown in **Fig S3C-D.** There were no significant correlations between the session-mean AUC dopamine release for each press type and the mean inter-press interval for that lever for that session (**Fig S3C**). Likewise, there was no significant correlation between the inter-event interval (i.e., seconds since preceding entry or press) and the AUC during the press for either First Presses or Mag→Presses (**Fig S3D**), however there was a significant negative relationship between IPSI Press→Press sequences and dopamine release, indicating that the inhibition of release was greater for longer Press→Press intervals during press sequences.

To examine the effect of a change in outcome identity on dopamine lateralization during the action window, we reversed the outcomes earned by each press mid-session such that each lever earned the opposite outcome on the same reinforcement schedule (**Fig 4A**, bottom). The $\Delta F/F$ signals during IPSI and CONTRA presses for one of the rats averaged across the pre-reversal and post reversal phases of the reversal session is shown in **Fig 4F** (left and right, respectively), and the mean press rates and dopamine AUCs for all rats during this session are presented in **Fig 4G**. There were no differences between IPSI and CONTRA press rates either pre-reversal or post-reversal (**Fig 4G** open bars, left and right panels, respectively, F's < 1.0). There was, however, a significant change in dopamine lateralization after reversal; whereas there was clear hemispheric lateralization during the pre-reversal stage (**Fig 4F**, left and **Fig 4G** filled bars, left panel IPSI vs CONTRA, k=4, p-critical=0.01, p=3.85E-06) this difference was abolished after reversal (**Fig 4F**, right and **Fig 4G**, filled bars, right panel, p=0.1). We subdivided

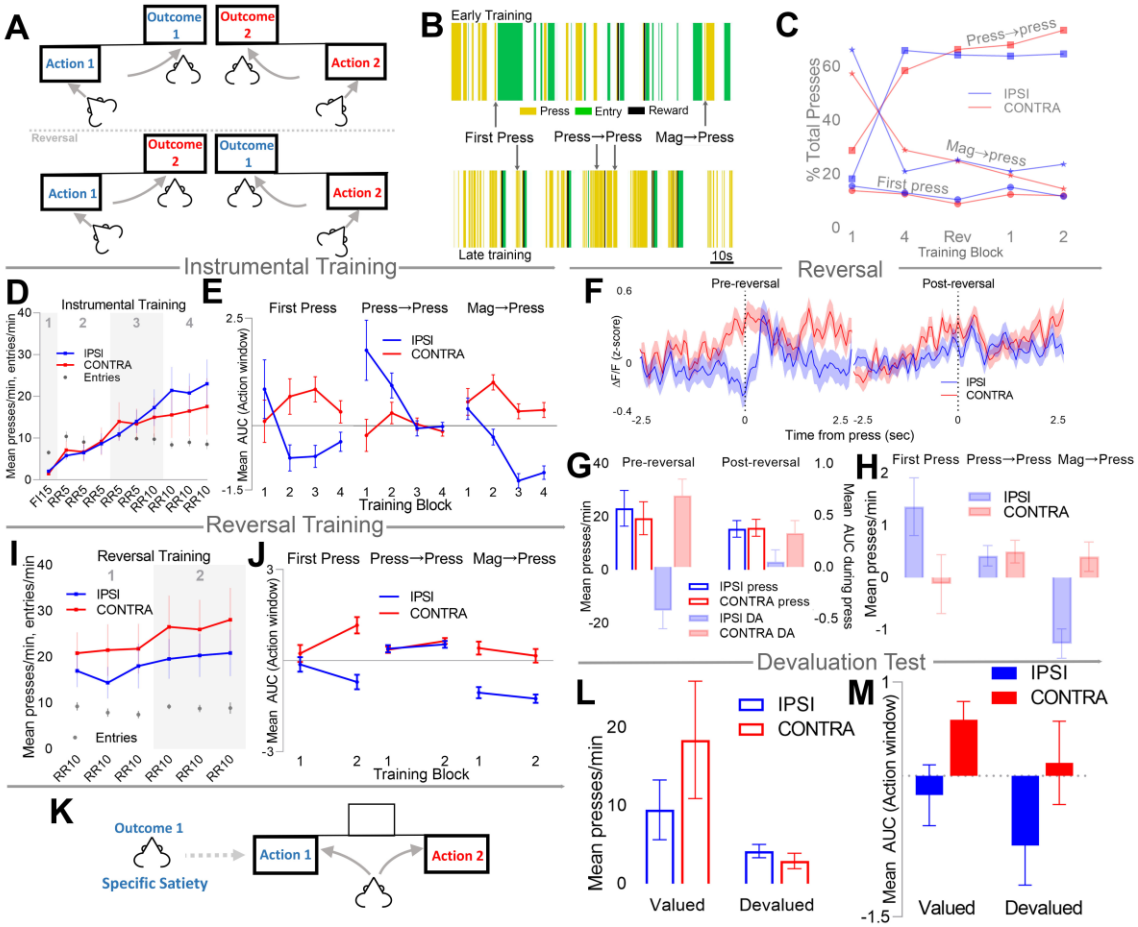

Fig. 4. The effect of reversing the outcome identity on dopamine release in DMS. (**A**) Rats were trained with two specific action-outcome pairings (instrumental training), the outcomes of which were then reversed (reversal), followed by several days of training with the new action-outcome pairings (reversal training). (**B**) Behavioral sequence over 2 min for one rat during Block 1 of instrumental training (top) and the same rat on Block 4 (bottom), lever presses in yellow, entries in green and rewards in black, press types identified. (**C**) Each press type represented as a percent of total presses on the first and last day of instrumental training training and reversal training, and on the day of reversal. (**D**) Mean (±SEM) presses per minute on the IPSI and CONTRA levers, and mean (±SEM) entries per minute (Entries) across 10 days of instrumental training. Numbers in grey indicate training blocks. (**E**) Mean (±SEM) AUC of the ΔF/F signals during the Action window for IPSI (blue) and CONTRA (red) press subtypes, averaged within each block of instrumental training. (**F**) ΔF/F signals averaged across IPSI (blue) and CONTRA (red) presses (indicated by dotted lines) for one rat during the reversal session before outcome reversal (left) and after outcome reversal (right). (**G**) Mean (±SEM) presses per minute (open bars) and AUC of the ΔF/F signal during the action window (filled bars) for IPSI (blue) and CONTRA (red) actions during the reversal session before reversal (left) and after reversal (right). (**H**) Mean (±SEM) AUC of the ΔF/F signal during the action window for press subtypes averaged across the reversal session after reversal. (**I**) Mean (±SEM) presses per minute on the IPSI and CONTRA levers, and mean (±SEM) entries per minute (Entries) across 6 days of reversal training. Numbers in grey indicate training blocks. (**J**) Mean (±SEM) AUC of the ΔF/F signals during the Action window for IPSI (blue) and CONTRA (red) press subtypes, averaged within each block of reversal training. (**K**) One outcome was devalued by allowing rats to feed on it to satiety, immediately followed by a choice test with two levers under extinction. (**L**) Mean (±SEM) presses per minute on the IPSI (blue) and CONTRA (red) levers when they were valued (left) and devalued (right). (**M**) Mean (±SEM) AUC of the ΔF/F signal during the Action window for IPSI and CONTRA presses when they were valued (left) and devalued (right).

the dopamine AUC signals according to press type during the reversal stage (**Fig 4H**); dopamine lateralization was abolished during First Presses (IPSI vs CONTRA First Press, p=0.8), however lateralization was preserved during Mag→Presses (IPSI vs CONTRA, k=3, p-critical=0.017, p=3.66E-05), and there remained no difference during Press→Press sequences (p=0.07). Therefore, outcome identity reversal abolished the hemispheric lateralization during press sequence initiations (First Presses).

To confirm that hemispheric lateralization re-develops across extended reversal training as animals learned the new action-outcome associations, we continued training rats for a further six days on the reversed contingencies, in an otherwise identical manner to initial training. Response rates across reversed training are presented in **Fig 4I**; there was a significant increase in press rates across days (linear trend F(1,5)=8.4, p=0.03), and no difference between IPSI and CONTRA press rates (F < 1.0). The mean dopamine AUCs during each press type are presented in **Fig 4J.** As in initial training, hemispheric lateralization during First Presses re-developed across the course of reversal training; there was no difference during training Block 1, but a significant difference by Block 2 (k=6, p-critical = 0.008, First Press IPSI vs CONTRA; Block 1 p=0.3, Block 2 p=3.35E-07). As in initial training and reversal, there remained no difference in dopamine release during IPSI and CONTRA Press→Press sequences (p's > 0.4), and the lateralization during Mag→Presses persisted throughout reversal training (p's < 2.68E-07). These data confirm, therefore, that continued training re-established the hemispheric lateralization during the initiation of action sequences.

Finally, we hypothesized that continued training allowed animals to update the goal-directed action-outcome association with the new outcome identities, and this was reflected in dopamine lateralization during First Presses. In order to test whether rats had indeed updated this relationship, we tested them for sensitivity to outcome devaluation using specific satiety(Balleine & Dickinson, 1998). Rats were given free access to one of the earned outcomes to satiety and were then given an extinction choice test with both levers available (**Fig 4K**). The press rates on the valued and devalued IPSI and CONTRA levers across the devaluation test are presented in **Fig 4L**. Rats showed a clear bias in performance (i.e. valued > devalued) in accordance with the reversed contingencies (main effect of value; Valued vs. Devalued F(1,5)=10.76, p=0.02), and there was no difference in overall press rates on each lever (no main effect of side; IPSI vs. CONTRA, F < 1.0). Dopamine release during this test (**Fig 4M**) reflected both the goal-directed action-outcome association and the overall outcome value: there was significant hemispheric lateralization (main effect of side; IPSI vs CONTRA, F(1,2150)=9.51, p=0.0021), and bilateral modulation according to the outcome value, where dopamine release during the action was lower when the outcome of the action was devalued (main effect of value; Valued vs Devalued, F(,2150)=4.6, p=0.032). Together, these results confirm that dopamine lateralization during actions reflects the goal-directed action-outcome association and suggest that a shift in net (i.e., bilateral) dopamine reflects the devaluation-induced change in action value.

Discussion

The decision to engage in an action to achieve a specific goal depends on integrating the causal relationship between the action and its outcome with the current value of that outcome(Balleine, 2019; Balleine & O'Doherty, 2010). Whereas the former reflects a stable association that gradually develops and changes only after extended experience, the latter reflects a dynamic state that can fluctuate on a trial-by-trial basis according to specific changes in the animal's state and the immediate reward history of the action (i.e. the action value). Here, we have shown that dopamine release in the pDMS during goal-directed actions simultaneously conveys these distinct streams of information relating to both the strength of the specific action-outcome association and action values. We found that trial-by-trial fluctuations in bilateral dopamine release during the action were associated with changes in action value and those changes during the outcome reported a reward prediction error. However, we found evidence of another signal that emerged with experience that broadcast the overall strength of the specific action-outcome relationship in the degree of lateralized dopamine release; i.e., in the difference between release in the hemisphere contralateral to the action and in the hemisphere ipsilateral to the action, during action sequence initiations. In this way, dopamine signaling conveys both the strength of the action-outcome relationship and the current value of the action, the integration of which lies at the heart of the acquisition and performance of goal-directed actions.

Difficulty reconciling the functions of striatal dopamine release in reward prediction, movement and motivation has provided a significant barrier to understanding its role in goal-directed action. A prevailing view has been that reward prediction errors are signaled in rapid 'phasic' dopamine, whereas slower variations in dopamine, as well as 'tonic' levels permit movement and motivation(Niv et al., 2007; Schultz, 2007). There has, however, been little or no evidence for such slower signals using techniques that allow fast timescale measurements(Mohebi et al., 2019). Furthermore, while there is evidence for anatomically distinct dopaminergic axonal signaling related to locomotion and reward(da Silva et al., 2018; Howe & Dombeck, 2016; Panigrahi et al., 2015), where these signals interact, it remains a challenge to discern how they are dissociated at the level of dopamine release; i.e., when increased dopamine release signals reward and when it signals action. Lateralized dopamine release during action performance has been interpreted as movement signals (R. S. Lee et al., 2019) (Parker et al., 2016), and indeed we saw some evidence of movement signals in actions that followed magazine entries in sequence. However, we were able to link the lateralized signal during sequence initiations directly to the strength of the action-outcome association and show that this signal was distinct from the lateralized movement signals observed during the subsequent reward approach and retrieval. Dopamine release was similar during ipsilateral and contralateral instrumental actions during initial training and diverged over time, whereas release during reward retrieval was lateralized from the outset of training. Furthermore, whereas dopamine lateralization during reward retrieval persisted when the action-outcome relationship was degraded, lateralization during sequence initiations was abolished during performance of the degraded action.

Not only was this signal sensitive to the action-outcome contingency, it was also sensitive to changes in the identity of the outcome; reversal of the outcome identities likewise abolished

lateralization which then re-developed as the new action-outcome associations were established. This finding together with the replication across schedules of reinforcement, provides consistent evidence that this lateralized signal reflects a learning-induced change that tracks the strength of the specific action-outcome association, necessary for the initiation of goal-directed action sequences, and indeed suggests a causal relationship between SNc dopamine neuron activity and the initiation of those sequences(da Silva et al., 2018). Furthermore, the finding that this lateralized signal emerged rapidly and progressively increased in magnitude across training reconciles traditional functions ascribed to slow changes in dopamine release associated with motivation or response vigor(da Silva et al., 2018; Niv et al., 2007; Panigrahi et al., 2015; Parkinson et al., 2002; Schultz, 2007) with more recent demonstrations that such changes, when measured on a faster timescale, are comprised of fast fluctuations in dopamine signaling(Mohebi et al., 2019).

The trial-by-trial changes in bilateral dopamine release tracks the updated action values. Previous reports have suggested that GCaMP activity in DMS-projecting dopamine neurons reflects action values(R. S. Lee et al., 2019; Parker et al., 2016), with dopamine signals in the ventral striatum encoding value and reward prediction-error(Mohebi et al., 2019). Importantly, however, we found that action value coding only occurred when an action's value was updated immediately following contact with reward or non-reward during outcome retrieval. There was substantial evidence to support this conclusion: We found no evidence for action value coding during press-press sequences but, more importantly, there was clear evidence of both positive and negative outcome-induced prediction-error coding by dopamine release during reward retrieval, confirming that a dopamine error signal was indeed generated at this time to allow the value of the action that immediately followed that retrieval to be updated. These results confirm the longstanding assumption that, as in Pavlovian conditioning, dopamine release during exposure to the instrumental outcome is bilaterally modulated by the reward prediction error and demonstrate that it is this error that updates the action values. In the final experiment, we demonstrated that shifts in outcome (and thus action) value with specific satiety (outcome devaluation), also engendered a shift in bilateral dopamine release, while maintaining hemispheric lateralization signaling the action-outcome association. Following an outcome devaluation treatment, animals showed clear hemispheric lateralization, in addition to a bilateral downward shift in dopamine release when animals were pressing the devalued lever. This example clearly shows how the action-outcome association and action value are integrated at the level of striatal dopamine, to allow goal-directed actions.

Materials and Methods

Animals

The experiments were conducted with healthy, experimentally naïve wild-type outbred Long-Evans rats aged between 8-16 weeks old prior to surgery (244-646g prior to surgery), obtained from the Randwick Rat Breeding Facility at the University of New South Wales. Rats were housed in transparent plastic boxes of 2-4 in a climate-controlled colony room and maintained on a 12 h light/dark cycle (lights on at 7:00 am). All experimental stages occurred during the light portion. Water and standard lab chow were continuously available prior to the

start of the experiment. All experimental and surgical procedures were approved by the Animal Ethics Committee at the University of New South Wales and are in accordance with the guidelines set out by the American Psychological Association for the treatment of animals in research.

Exclusions and counterbalancing:
Contingency Degradation
Subjects were 30 rats (13 female). All training and lever allocations were fully counterbalanced between males and females, left and right hemisphere recordings and within housing boxes with regards to order of lever and outcome presentation and identity of lever and outcome being degraded. Following histological assessment for placement of dLight1.1 and fiber optic in the pDMS, there were 9 rats excluded for misplaced cannulas or virus injections, leaving a total of 21 rats in the experiment (10 female), with 6 right pDMS recordings (3 ipsilateral lever degraded) and 15 left pDMS recordings (7 ipsilateral lever degraded).

Outcome Identity Reversal
Subjects were 8 rats (4 female). All training and lever allocations were fully counterbalanced as in the Contingency Degradation experiment except that all recordings were made in the left hemisphere. Two rats were excluded from the study due to misplaced fiber optic cannulae, leaving a total of 6 rats in the experiment (3 female).

Surgery

Rats were anaesthetized with 3% inhalant isoflurane gas with oxygen, delivered at a rate of 0.5L/min throughout surgery. Anaesthetized rats were placed in a stereotaxic frame (Kopf), and an incision was made down the midline of the skull, and the scalp was retracted to expose the skull. A 1.0 uL glass Hamilton syringe was lowered into the brain for infusions of dLight1.1 (pAAV5-CAG-dLight1.1, AddGene #111067-AAV5), which was delivered at a total volume of 0.75uL (0.1uL/min), with the syringe left in place for an additional 3 min to allow for diffusion. A fiber optic cannula (400um, 6mm, Doric) was then implanted above the injection site, and secured in place with dental cement, attached to the skull with 3 jeweller's screws. Following surgery, rats were injected with a prophylactic (0.4 mL) dose of 300 mg/kg procaine penicillin. Rats were given a minimum of 3 weeks of recovery time following surgery to allow sufficient viral expression.

Surgical co-ordinates were pre-determined from pilot studies and varied slightly between males and females, with dLight infused into males at the co-ordinates (mm from Bregma) A/P:-0.5, M/L:$\pm$2.6, D/V:-4.8 and for females A/P:-0.4, M/L:$\pm$2.55, D/V:-4.8, with fiber optic placement targeted 0.1mm dorsal to these injection sites.

Apparatus

Training was conducted in two MED Associates operant chambers enclosed in sound- and light-attenuating cabinets. Each chamber was fitted with a pellet dispenser capable of delivering a 45 mg grain food pellet (Bioserve Biotechnologies), as well a pump fitted with a syringe outside the chamber, capable of delivering 0.2 mL of 20% sucrose solution (white sugar, Coles,

Australia) diluted in water, each delivered to separate compartments of a recessed magazine inside the chamber.

The chambers also contained two retractable levers that could be inserted individually on the left and right sides of the magazine. Head entries into the magazine were detected via an infrared photobeam. Unless otherwise stated, the operant chambers were fully illuminated during all experimental stages by a 3W, 24V house light located on the upper edge of the wall opposite to the magazine. All training sessions were pre-programmed on a computer through the MED Associates software (Med-PC).

Fiber Photometry

Fiber photometry recordings were conducted using a dedicated fiber photometry processor and software (RZ5P Processor, Synapse, Tucker-Davis Technologies, TDT), which were used to control LEDs (465nm excitation and 405nm isosbestic) via LED drivers with a Fluorescence Mini Cube, measured with Newport photoreceivers, all from Doric Lenses. Patch cords used for recordings (400um, Doric) were photobleached for a minimum of 45mins at the start of each recording day. Light was measured at the tip of the patch prior to recording sessions, and maintained at 10-20uW (for 465) and 2-7uW (for 205). LEDs were modulated and demodulated via Synapse software at 331 Hz (465) and 211 Hz (405) and low-pass filtered (6Hz). MedPC signals were sent to RZ5P/Synapse, to indicate reward deliveries, magazine entries and lever presses, which were timestamped into the fiber photometry recordings.

Behavioral Protocol and Food Restriction

Food restriction (all experiments). Rats underwent 2-4 days of food restriction prior to the onset of magazine training and this continued throughout the duration of the experiment. During this time, they received 5 g each of chow daily in addition to 1 g each of grain pellets and 2hr/daily exposures to sucrose solution. Rats were then given 10-20 g of chow daily from the fourth day until the end of the experiment. Their weight was monitored to ensure it remained above 85% of their pre-surgery body weight at all times.

Magazine training (all experiments). Rats were given two days of magazine training prior to the onset of instrumental conditioning, which consisted of 15 deliveries each of pellets and sucrose into the magazine, delivered intermingled (but non-overlapping) at random intervals of 60 s.

Contingency Degradation Experiment

Instrumental training. Rats were trained on an alternating lever paradigm, in which each lever came out for a total of 15 minutes or 10 outcomes, whichever came first, in alternating fashion, such that each lever was presented twice (total of 20 outcomes available per lever). On Day 1 of instrumental training (Block 1), rats were trained to press both levers on a fixed interval 15-sec schedule (FI15), where reinforcements were separated by a fixed interval of minimum 15 seconds, but every response spaced by more than 15 seconds was reinforced. This continued until rats reached criterion of earning 40 outcomes in a single session.

Over the following 3 days (Days 2-4 of instrumental training, Block 2), rats were trained on an RI15 sec schedule (presses were reinforced on average every 15 secs), on the same alternating levers paradigm, with outcome criterion increased to 30 of each outcome in each

session (15 outcomes per lever presentation). Over the following 6 days (Days 5-10, Blocks 3-4), rats were trained on the same paradigm on an RI30 sec schedule (presses reinforced on average every 30 seconds).

Contingency degradation. Over the following 10 days (5 x 2-day blocks), rats were trained on a contingency degradation schedule, in which rewards were earned on the same RI30 schedule with alternating levers, with one outcome (pellets or sucrose) additionally delivered non-contingently on an RI45 sec schedule, explicitly unpaired with instrumental responses or magazine entries (no free outcomes could be delivered within 1 sec of a lever press or magazine entry), delivered during both lever presentations, but not during the 1min interval between levers. This schedule was chosen to approximately match the number of free and earned outcomes of the same type (rats received 31 free outcomes on average at the start of training and 36 on average by the end of training).

Test. On the final day after contingency degradation, rats were given a choice test under extinction (i.e. no outcomes were delivered) with both levers for 20 minutes. Rats were placed in the chambers, and after a 1 min interval, the house lights were illuminated and both levers came out for 20 mins, after which the session ended.

Outcome Identity Reversal Experiment
Instrumental Training. Lever press training was administered as in the Contingency Degradation experiment with the following differences: after achieving criterion on the FI15 schedule (Block 1; 15 minutes or 10 outcomes per lever presentation), rats were trained on a Random Ratio-5 (RR5) schedule for 5 days (Block 2) where outcomes were earned after an average of 5 responses (15 minutes or 15 outcomes per lever presentation for the remainder of the experiment), followed by 4 sessions of RR10 training (Block 3).

Mid-Session Reversal. Maintaining the RR10 schedule of reinforcement, animals underwent a single session where the outcomes paired with each action were reversed mid-session. This session had 6 alternating lever presentations (order counterbalanced) where the initial action-outcome pairings remained intact for the first 2 presentations (i.e. R1-O1, R2-O2) and were reversed for the remaining 4 (i.e. R1-O2, R2-O1).

Reversal Training. Rats were trained on the reversed contingencies (i.e. R1-O2, R2-O1) for a further 6 sessions (RR10) as in Instrumental Training.

Outcome Devaluation Testing. All rats received two 1-hour pre-exposure sessions to their individual devaluation chambers prior to the outcome devaluation tests. For outcome devaluation tests, rats were placed in their devaluation chamber for 30 minutes with *ad libitum* access to either grain pellets or sucrose solution (counterbalanced for pairing with lever relative to recording hemisphere). Rats were then immediately placed in the experimental chambers where they were given a 5-minute choice test in extinction with both levers available. This test was conducted twice for each rat (after devaluation of each outcome separately) with a session of re-training on the intervening day with the reversed contingencies maintained. Behavioural

and fiber photometry data for one rat was excluded for one outcome devaluation test due to a lack of consumption of the outcome during devaluation pre-feeding (<1g).

Histology and Immunofluorescence Staining

Within a week after the final contingency degradation or devaluation tests, the rats were perfused with 4% paraformaldehyde in phosphate buffer (PB), and the brains sliced coronally on a vibratome at 40μm. A minimum of 6 sections each from the pDMS (from +1.2mm to -0.6mm A/P from Bregma) were stained for dLight1.1 and DARP-32. Sections were rinsed 3 times for 10 min each in 0.1 M phosphate buffered saline (PBS), then submerged in a blocking solution of PBS with 0.5% Triton and 10% normal horse serum (NHS) for 2 hrs at room temperature. Sections were then submerged in 600 μL of rabbit anti-GFP (1:1000, Invitrogen) and mouse anti-DARP-32 (1:1000, BD Biosciences) diluted in PBS with 0.2% Triton and 2% NHS for 48 hrs at 4 degrees C. Sections were then rinsed 3 times for 10 min in PBS, before being submerged in donkey-anti-rabbit Alexa 488 (1:1000, Invitrogen) and donkey-anti-mouse Alexa 546 (1:1000, Invitrogen), diluted in PBS with 0.2% Triton and 2% NHS for 2 hrs at room temperature. Sections were washed twice with PBS and twice with PB for 10 min each, and mounted with Vectashield mounting medium with DAPI (Vector Laboratories).

Placement of dLight1.1 injections and fiber cannulas were imaged 24-48 hours later under a confocal microscope (BX16WI, Olympus) with a 10x objective using the boundaries defined by Paxinos & Watson (2016). Rats were excluded from the analysis if the placement of the fiberoptic cannula or virus extended anterior to +0.2mm from bregma in the DMS, or cannula placement was outside of the boundaries of the pDMS.

Fiber Photometry Analysis

Fiber photometry data were analyzed in custom MATLAB scripts. To achieve a motion-artifact corrected dLight signal, the demodulated dLight signal (465nm) and isosbestic control signal (405nm) were extracted from the recording software (Synapse), and the isosbestic signal was fitted to the dLight signal using a least-squares linear fit model for each recording session. Change in fluorescence ($\Delta F/F$) was calculated as (dLight signal – fitted isosbestic)/fitted isosbestic. Peri-event $\Delta F/F$ signals were extracted for left lever presses, right lever presses and magazine entries, across a 10s window (5s before and 5s after event onset) and downsampled to 15.8 samples per second (i.e. non-overlapping windows of 64 samples were averaged into a single value). The $\Delta F/F$ signal was normalized within these windows to the first 2.5s of each 10s window (baseline period), to give z-scores indicating the mean change from baseline ($\Delta F/F$ minus the mean baseline $\Delta F/F$ divided by the standard deviation of the baseline $\Delta F/F$).

For event-related analysis, the area under the curve (AUC; calculated via the trapezoidal method for integral calculation; trapz function in MATLAB) was calculated for each event across event windows: Action: 0.2s before to 0.2s after press; and Outcome: 0 to 1.5s after magazine entry. These AUCs were averaged to give mean AUCs for each event type.

To examine the relationship between reward prediction error (δ) and the temporal dynamics of the dopamine release after magazine entry, we employed a bootstrapping confidence intervals procedure (bCI; *37*) on dLight signals aligned to the magazine entry for high δ vs low δ CONTRA-Rew, CONTRA-Empty, IPSI-Rew and IPSI-Empty entries. This approach allows for the direct comparison of waveform profiles for two groups of signals (high δ vs low δ within

each category). For a given event category, the population of Z-scored ΔF/F signals were randomly resampled with replacement for the same number of trials in that category and the mean waveform calculated. This was repeated 1000 times (i.e. 1000 bootstraps) to create a distribution of bootstrapped means. The differences between mean resampled waveforms (i.e. high δ - low δ) were calculated at each timepoint, and 95% CIs were then calculated for each timepoint in the series (taking the 2.5 and 97.5 centiles from the ranked bootstrap distribution, expanded by $\sqrt{(n/(n-1))}$ to correct for the narrowness bias). CIs that did not contain zero were identified, with at least two such consecutive timepoints constituting a significant difference (imposed to minimize the detection of spurious transients and corresponding to a consecutive threshold period of ~0.13s).

Reinforcement Learning

Reinforcement learning was modelled using a simplified temporal difference learning rule, in which the action value, V, was updated according to an error term (δ) generated during magazine entry according to $\delta(t)=\alpha[R(t)-V(t)]$, whereby α was a learning rate parameter and R was set at 1 for rewarded entries and 0 for nonrewarded entries. The action value of the next action, V(t+1) was updated according to the sum of the preceding action value and its error $V(t+1) = V(t) + \delta(t)$. Action values were calculated separately for each instrumental action (IPSI and CONTRA), all were set at 0 for the very first session, and for all subsequent sessions, V-values started at the final value of V for that action of the preceding session. The learning rate (α), was set at 0.3, however the pattern of data was preserved when learning rates were varied over a range of values (from 0.1-0.5) (**Fig S2C**).

There was a modification to this model that was made to separately assess the possibility of action values being updated on every press, which allowed for negative prediction errors in the absence of magazine entries, whereby δ was calculated to generate a negative term following all presses in which there wasn't a rewarded magazine entry, and V was updated on every press.

Signal data was compiled across the Action window according to the value of V at the time of press: 0-0.3 (low); 0.3-0.6(low-med); 0.6-0.8(med-high) and 0.8-1(high). The same was done for δ during Outcome window: 0-.075(low); 0.075-0.15(low-med);0.15-0.23(med-high) and 0.23-0.3(high), with the maximum error term being equivalent to α.

Statistical Analysis

For all analyses, non-orthogonal comparisons controlled the Type I error rate at α/k, for k contrasts, with $\alpha=0.05$, according to the Bonferroni correction. For planned orthogonal comparisons, significance was set at $\alpha=0.05$.

Behavioral data were analyzed according to planned contrasts comparing response rates on the degraded (or to-be-degraded) and non-degraded lever across instrumental training, contingency degradation training, and on test. Instrumental response rates during contingency degradation were converted into baseline scores, in order to account for pre-existing differences in response rates on each lever. This was taken as the response rate on each lever represented as a percentage of the mean response rate on that same lever across the final block of instrumental training (i.e. Block 4, Days 8-10).

For reversal training, orthogonal contrasts tested for differences in responding on the IPSI and CONTRA levers across training, reversal and reversal training.

For correlation analyses between AUC and inter-press and inter-event intervals, correlations were analyzed using simple linear regression, with Bonferroni-corrected p-values adjusted according to p(adjusted) = p*k, (k=12) for 12 comparisons (4 training blocks x 3 press types each for IPSI and CONTRA). Significance was set at $p<0.05$.

For analysis of fiber photometry AUCs across training (all experiments), pairwise comparisons (t-tests) were planned to test the difference between IPSI and CONTRA presses according to press type, within each training block, and averaged across all training blocks. During entries, tests compared the mean AUC during rewarded and empty magazine entries that followed either IPSI or CONTRA presses, within each training block. For assessment of prediction-error modulation of dopamine, planned comparisons tested the magnitude of dopamine within each press type at the highest and lowest V or δ values. For contingency degradation, comparisons were planned to test the difference between degraded and non-degraded presses within each press type. For mid-session reversal, comparisons were planned to test the difference between IPSI and CONTRA presses, averaged across all press types, before and after reversal, and within each press type after reversal. For all experiments, significance was set at $p<\alpha/k$ (Bonferroni).

For outcome devaluation, behavioral and fiber photometry data were analyzed using planned, orthogonal contrasts testing for a main effect of side (IPSI vs CONTRA) and a main effect of value (Valued vs Devalued) controlling α at 0.05.

Data availability

All source data are published in Figshare https://figshare.com/articles/dataset/Striatal_dopamine_encodes_the_relationship_between_actions_and_rewards/19083647. dLight1.1 was obtained under material transfer agreement with AddGene. All other materials for fiber photometry recordings were obtained from Tucker-Davis Technologies and Doric Lenses.

References and Notes

Balleine, B. W. (2019). The Meaning of Behavior: Discriminating Reflex and Volition in the Brain. *Neuron*, *104*(1), 47–62. https://doi.org/10.1016/j.neuron.2019.09.024

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4–5), 407–419.

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, *35*(1), 48–69. https://doi.org/10.1038/npp.2009.131

Balleine, B. W., Peak, J., Matamales, M., Bertran-Gonzalez, J., & Hart, G. (2021). The dorsomedial striatum: An optimal cellular environment for encoding and updating goal-directed learning. *Current Opinion in Behavioral Sciences*, *41*, 38–44. https://doi.org/10.1016/j.cobeha.2021.03.004

Bradfield, L. A., Bertran-Gonzalez, J., Chieng, B., & Balleine, B. W. (2013). The thalamostriatal pathway and cholinergic control of goal-directed action: Interlacing new with existing learning in the striatum. *Neuron*, *79*(1), 153–166. https://doi.org/10.1016/j.neuron.2013.04.039

Colwill, R. M., & Rescorla, R. A. (1986). Associative Structures In Instrumental Learning. In *Psychology of Learning and Motivation* (Vol. 20, pp. 55–104). Elsevier. https://linkinghub.elsevier.com/retrieve/pii/S007974210860016X

da Silva, J. A., Tecuapetla, F., Paixão, V., & Costa, R. M. (2018). Dopamine neuron activity before action initiation gates and invigorates future movements. *Nature*, *554*(7691), 244–248. https://doi.org/10.1038/nature25457

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*(1), 1–18. https://doi.org/10.3758/BF03199951

Dickinson, A., & Charnock, D. J. (1985). Contingency Effects with Maintained Instrumental Reinforcement. *The Quarterly Journal of Experimental Psychology Section B*, *37*(4b), 397–416. https://doi.org/10.1080/14640748508401177

Dolan, R. J., & Dayan, P. (2013). Goals and Habits in the Brain. *Neuron*, *80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Fisher, S. D., Ferguson, L. A., Bertran-Gonzalez, J., & Balleine, B. W. (2020). Amygdala-Cortical Control of Striatal Plasticity Drives the Acquisition of Goal-Directed Action. *Current Biology*, *30*(22), 4541-4546.e5. https://doi.org/10.1016/j.cub.2020.08.090

Gerfen, C. R., & Surmeier, D. J. (2011). Modulation of striatal projection systems by dopamine. *Annual Review of Neuroscience*, *34*, 441–466. https://doi.org/10.1146/annurev-neuro-061010-113641

Hart, G., Bradfield, L. A., & Balleine, B. W. (2018). Prefrontal Corticostriatal Disconnection Blocks the Acquisition of Goal-Directed Action. *The Journal of Neuroscience: The Official Journal*

of the Society for Neuroscience, *38*(5), 1311–1322. https://doi.org/10.1523/JNEUROSCI.2850-17.2017

Hart, G., Bradfield, L. A., Fok, S. Y., Chieng, B., & Balleine, B. W. (2018). The Bilateral Prefronto-striatal Pathway Is Necessary for Learning New Goal-Directed Actions. *Current Biology: CB*, *28*(14), 2218-2229.e7. https://doi.org/10.1016/j.cub.2018.05.028

Holly, E. N., Davatolhagh, M. F., Choi, K., Alabi, O. O., Vargas Cifuentes, L., & Fuccillo, M. V. (2019). Striatal Low-Threshold Spiking Interneurons Regulate Goal-Directed Learning. *Neuron*. https://doi.org/10.1016/j.neuron.2019.04.016

Howe, M. W., & Dombeck, D. A. (2016). Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature*, *535*(7613), 505–510. https://doi.org/10.1038/nature18942

Klaus, A., Martins, G. J., Paixao, V. B., Zhou, P., Paninski, L., & Costa, R. M. (2017). The Spatiotemporal Organization of the Striatum Encodes Action Space. *Neuron*, *95*(5), 1171-1180.e7. https://doi.org/10.1016/j.neuron.2017.08.015

Lee, J., Wang, W., & Sabatini, B. L. (2020). Anatomically segregated basal ganglia pathways allow parallel behavioral modulation. *Nature Neuroscience*, *23*(11), 1388–1398. https://doi.org/10.1038/s41593-020-00712-5

Lee, R. S., Mattar, M. G., Parker, N. F., Witten, I. B., & Daw, N. D. (2019). Reward prediction error does not explain movement selectivity in DMS-projecting dopamine neurons. *ELife*, *8*. https://doi.org/10.7554/eLife.42992

Matamales, M., McGovern, A. E., Mi, J. D., Mazzone, S. B., Balleine, B. W., & Bertran-Gonzalez, J. (2020). Local D2- to D1-neuron transmodulation updates goal-directed learning in the striatum. *Science (New York, N.Y.)*, *367*(6477), 549–555. https://doi.org/10.1126/science.aaz5751

Mohebi, A., Pettibone, J. R., Hamid, A. A., Wong, J.-M. T., Vinson, L. T., Patriarchi, T., Tian, L., Kennedy, R. T., & Berke, J. D. (2019). Dissociable dopamine dynamics for learning and motivation. *Nature*, *570*(7759), 65–70. https://doi.org/10.1038/s41586-019-1235-y

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology*, *191*(3), 507–520. https://doi.org/10.1007/s00213-006-0502-4

Nonomura, S., Nishizawa, K., Sakai, Y., Kawaguchi, Y., Kato, S., Uchigashima, M., Watanabe, M., Yamanaka, K., Enomoto, K., Chiken, S., Sano, H., Soma, S., Yoshida, J., Samejima, K., Ogawa, M., Kobayashi, K., Nambu, A., Isomura, Y., & Kimura, M. (2018). Monitoring and Updating of Action Selection for Goal-Directed Behavior through the Striatal Direct and Indirect Pathways. *Neuron*, *99*(6), 1302-1314.e5. https://doi.org/10.1016/j.neuron.2018.08.002

Panigrahi, B., Martin, K. A., Li, Y., Graves, A. R., Vollmer, A., Olson, L., Mensh, B. D., Karpova, A. Y., & Dudman, J. T. (2015). Dopamine Is Required for the Neural Representation and

Control of Movement Vigor. *Cell*, *162*(6), 1418–1430. https://doi.org/10.1016/j.cell.2015.08.014

Park, J., Coddington, L. T., & Dudman, J. T. (2020). Basal Ganglia Circuits for Action Specification. *Annual Review of Neuroscience*, *43*(1), 485–507. https://doi.org/10.1146/annurev-neuro-070918-050452

Parker, N. F., Cameron, C. M., Taliaferro, J. P., Lee, J., Choi, J. Y., Davidson, T. J., Daw, N. D., & Witten, I. B. (2016). Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nature Neuroscience*, *19*(6), 845–854. https://doi.org/10.1038/nn.4287

Parkinson, J. A., Dalley, J. W., Cardinal, R. N., Bamford, A., Fehnert, B., Lachenal, G., Rudarakanchana, N., Halkerston, K. M., Robbins, T. W., & Everitt, B. J. (2002). Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive Pavlovian approach behaviour: Implications for mesoaccumbens dopamine function. *Behavioural Brain Research*, *137*(1–2), 149–163. https://doi.org/10.1016/S0166-4328(02)00291-7

Patriarchi, T., Cho, J. R., Merten, K., Howe, M. W., Marley, A., Xiong, W.-H., Folk, R. W., Broussard, G. J., Liang, R., Jang, M. J., Zhong, H., Dombeck, D., von Zastrow, M., Nimmerjahn, A., Gradinaru, V., Williams, J. T., & Tian, L. (2018). Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. *Science*, *360*(6396), eaat4422. https://doi.org/10.1126/science.aat4422

Peak, J., Chieng, B., Hart, G., & Balleine, B. W. (2020). Striatal direct and indirect pathway neurons differentially control the encoding and updating of goal-directed learning. *ELife*, *9*, e58544. https://doi.org/10.7554/eLife.58544

Peters, A. J., Fabre, J. M. J., Steinmetz, N. A., Harris, K. D., & Carandini, M. (2021). Striatal activity topographically reflects cortical activity. *Nature*, *591*(7850), 420–425. https://doi.org/10.1038/s41586-020-03166-8

Rescorla, R. A. (1992). Response-independent outcome presentation can leave instrumental R-O associations intact. *Animal Learning & Behavior*, *20*(2), 104–111. https://doi.org/10.3758/BF03200407

Reynolds, J. N., Hyland, B. I., & Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*(6851), 67–70. https://doi.org/10.1038/35092560

Santos, F. J., Oliveira, R. F., Jin, X., & Costa, R. M. (2015). Corticostriatal dynamics encode the refinement of specific behavioral variability during skill learning. *ELife*, *4*, e09423. https://doi.org/10.7554/eLife.09423

Schultz, W. (2007). Multiple Dopamine Functions at Different Time Courses. *Annual Review of Neuroscience*, *30*(1), 259–288. https://doi.org/10.1146/annurev.neuro.28.061604.135722

Shiflett, M. W., & Balleine, B. W. (2011). Molecular substrates of action control in cortico-striatal circuits. *Progress in Neurobiology*, *95*(1), 1–13. https://doi.org/10.1016/j.pneurobio.2011.05.007

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

P. Jean-Richard-dit-Bressel, C.W.G. Clifford, G. McNally. Analyzing event-related transients: confidence intervals, permutation tests, and consecutive thresholds. *Front. Mol. Neurosc*i. Feb; 13:14 (2020).

Acknowledgments: The authors thank Amir Dezfouli for helpful comments on the manuscript, Myles Billard from Tucker-Davis Technologies for technical assistance with the fiber photometry recordings, and Karly Turner for assistance in the lab.

Funding: Funding for this research was provided by a Discovery Project Grant from the Australian Research Council: DP160105070 (BWB) and by a Senior Investigator Grant from the National Health and Medical Research Council of Australia # GNT1175420 (BWB)

Author contributions:

Conceptualization: GH, TB, BWB

Methodology: GH, TB, BWB

Investigation: GH, TB

Funding acquisition: BWB

Writing: GH, TB, BWB

Competing interests: Authors declare that they have no competing interests.

Materials and Correspondence: All correspondence and materials requests should be directed to Bernard Balleine bernard.balleine@unsw.edu.au.