Title: Dopamine reveals adaptive learning of actions representation

Authors: Maxime Come[1], Aylin Gulmez[1], Loussineh Keshishian[1], Joachim Jehl[1], Elise Bousseyrol[1], Steve Didienne[1], Eleonore Vicq[1], Tinaïg Le Borgne[1], Alexandre Mourot[1], Philippe Faure[1]*

Affiliations:

1. Brain Plasticity Laboratory, CNRS UMR 8249, ESPCI Paris, PSL Research University, Paris, France.

*Corresponding author: phfaure@gmail.com

Abstract:

Efficient decision-making requires two key processes: learning values from actions and identifying a set of relevant actions to learn from in a given context. While dopamine (DA) is a well-known substrate for signaling reward prediction errors (RPEs) from selected actions to adjust behavior, the process of establishing and switching between action representations is still poorly understood. To address this gap, we used fiber photometry and computational modelling in a three-armed bandit task where mice learned to seek rewards delivered through three successive rule sets, displaying distinct strategies in each rule. We show that DA dynamically reflected RPEs computed from different task features, revealing context-specific internal representations. Our findings demonstrate that mice not only learned and updated action values but also action representations, adapting the features from which they learn across rules for flexible adjustment of their decision strategy.

Introduction

Toddlers solving puzzles can successfully associate either shapes or colors depending on the game they are playing (**Fig 1A**), highlighting the importance of context in learning value from environmental featuress, and thereby developing an internal model of a task structure. Efficient decision making indeed requires both to learn from the consequences of actions (reinforcement learning) and to identify features and dimensions (i.e., a state space) that define a set of relevant actions from which to learn about (representation learning) (*1–4*). A cornerstone of understanding the mechanisms governing reinforcement learning and decision making is the interplay between prediction errors and state representation. Failure in such representation learning can lead to superstitions or false beliefs that interfere with efficient learning and decision making (*5*). Despite its fundamental importance for adaptive behavior, the role of representation learning in decision-making has been experimentally overlooked, limiting our understanding of how state representations are formed through experience (*4*). This issue becomes increasingly important as researchers shift their focus from experiments with a simple task structure to more elaborated tasks (*6–10*) that more closely resemble natural decision-making, with multiple (and possibly overlapping or competing) features that animals may use as state representations, as well as potentially abrupt changes over time in the state representations being used.

The identification of the neural substrate of this representation can be an indication that this representation is actually being used by the animal. While multiple brain areas contribute to the encoding of such features (*11–14*), it is still difficult to know, in a given context, which one of these features are recognized and effectively used by a subject to build a relevant internal model of the world, e.g., to predict values, compute errors, and guide goal-directed actions. We hypothesize that dopamine (DA) could be an excellent indicator of the representations used to navigate an environment. DA is a very well-established substrate to signal value and compute reward prediction error (RPE) (*15–26*), integrating outcome-related dimensions in a common currency (*27–29*), and driving reinforcement learning and decision making (*21*, *27*, *28*, *30–33*). Consequently, DA-mediated RPE should necessarily depend on the most relevant features for obtaining rewards and driving strategy, thereby providing insights into the subject's current state representation. To demonstrate this, we propose a novel experimental approach designed to follow the learning and shifts in task representations. We used behavioral assays, fiber photometry recording and computational modeling to explore how dopamine-mediated RPE signatures are related to specific features or action in different rules of a spatial bandit task (*7*, *8*, *34*, *35*) and how these features vary across rules. Our results show that mice not only learned value from actions, but also adapted their set of relevant actions from which to learn, efficiently adjusting their reward-seeking strategies.

Each reward context is associated with a specific reward-seeking strategy

Using different versions of a spatial bandit task adapted for mice (*7*, *8*, *34*, *35*), we aimed to obtain rule-specific and feature-dependent strategies (**Fig 1A-B**). In this task, animals learned to navigate between three marked locations in an open field, each associated with an intracranial self-stimulation (ICSS) of the medial forebrain bundle (MFB). Mice could not receive two consecutive ICSS at the same location; and therefore, had to alternate between rewarding locations, resulting in a sequence of movements and binary choices (i.e., trials) **(Fig 1B, top)**. Despite the apparent simplicity of this self-generated, goal-oriented behavior, mice can use different features of the environment to guide their actions and obtain rewards (**Fig 1B, bottom**). Mice were initially trained in a deterministic context (*Det*) where all locations consistently delivered ICSS, developing typical ballistic speed profiles **(Fig 1C)** and increasing trial numbers, with similar learning curves observed in both males and females **(Fig S1A)**. Subsequently, mice were switched to complex and probabilistic reward delivery rules, requiring them to adapt their strategies **(Fig 1D, Fig S1B-C)**. In the complex context (*Cplx*), reward delivery was determined by the variability compared to decision patterns identified in the previous nine choices **(Fig S2A)** (*8*), while the probabilistic context (*Proba*) offered different reward probabilities at each location (100%, 50%, 25%)(*35*). These varying conditions resulted in distinct trajectory patterns **(Fig 1D)**, success rates **(Fig 1E)**, and decision-making strategies. In *Det*, animals tended to adopt circular trajectories with minimal U-turns (~20%). In contrast, the *Cplx* rule resulted in random trajectory patterns characterized by high sequence complexity (**Fig 1E, Fig S2B)**. In *Proba*, mice exhibited a bias toward locations with higher probability of reward delivery, resulting in a high percentage of U-turns and a preference for p100 and p50 (**Fig 1F**). We also ensured that those differences in decision strategy were not due to motivation or vigor to perform the three versions of the task **(Fig S3)**. Overall, while the basic design of the task remained constant, each rule is associated with a specific reward structure promoting different action-outcome causalities. The evolution of decision dynamics across rules demonstrates that mice can extract such contingencies to dynamically adjust and improve their reward-seeking strategies, allowing for the longitudinal study of both choice behavior adaptations and their neural correlates.

Dopamine dynamics reveal expectations built upon rule-specific features

We next examined DA release dynamics during the task, across the three rules, using the fluorescent sensor GRAB$_{DA2M}$ expressed in the lateral shell of the nucleus accumbens (NAc) in a new cohort of wild-type male mice (**Fig 2A, Fig S4A**). Positive transients in DA release occurred upon receiving expected rewards, whereas negative events were observed when expected rewards were omitted **(Fig 2B-C, Fig S4B)**, indicative of a negative RPE (for simplicity, these events, whether positive or negative, are referred to as transients). Similar responses were observed while recording Ventral Tegmental Area (VTA) DA neurons activity with GCaMP in DAT-iCre mice, ensuring consistency in the interpretation of DA dynamics between release and firing processes **(Fig S4C-D)**. Analysis of the amplitude distribution of DA transients (positive and negative) across the different rules showed greater variability compared to unexpected random stimulation in a rest cage, suggesting an active mechanism related to reward expectation modulating the DA response, rather than being a mere response to the ICSS (**Fig 2D**). Additional experiments with unexpected rewards delivered either during the task but off-target (i.e. when the animal was in-between rewarded locations, **Fig 2E**) or in a rest cage (**Fig S4E-F**) demonstrated a larger transient compared to expected rewards during the task, yet only after conditioning **(Supp 4G-H)**, further supporting the role of expectation in modulating DA release. We also controlled for a potential impact of sensor fatigue and found no effect on DA signal when stimulations were given in the rest cage with varying durations in-between stimulation (matching those observed in the task, typically from 2s to 7s) **(Fig S5A-B)**. Altogether, these findings, consistent with positive and negative RPE patterns, illustrate that DA dynamics during the task are not solely driven by MFB stimulation but are significantly influenced by the mice's learned expectations and internal task representations.

We next wondered which task features those expectations were built upon. To do this, we applied generalized linear models (GLMs) to analyze fluctuations in DA peaks and dips amplitudes across trials, running separate regression analyses for each individual mouse at the end of each rule (last two sessions) **(Fig 2F)**. The predictors included current and previous trial outcomes (reward or omission), the specific target where outcomes occurred (locations pA, pB, and pC; or p100, p50 and p25 in Proba), and the direction taken (Forward movement or U-turn) **(Fig 2F)**. In the *Det* setting, where all trials were rewarded, we observed that the key predictor for differentiating trials was direction but not target **(Fig 2G).** In the *Cplx* setting, trial outcome accounted for the biggest part of DA variation (positive for rewards, negative for omissions, **Fig 2H**), with an additional positive effect of previous outcome (having received an omission at trial n-1 increases DA signal at trial n), regardless of targets or directions. In *Proba*, this effect of previous outcome disappeared, and the target probability significantly influenced DA variations **(Fig 2I)**.

Overall, the GLM analysis revealed that the primary drivers of DA fluctuations varied depending on the task setting, with direction, trial outcome, and target probability each playing distinct roles. Direct examinations of DA transients, categorized by direction, previous outcome or target, confirmed and complemented these results. In *Det*, DA release depends on direction **(Fig 2J, Fig S5C)** but not on the target (**Fig S5D**). In *Cplx*, omission on previous trial led to greater rewards-induced peaks and shallower omissions-induced dips (**Fig 2K, Fig S5E**), while neither the target nor the direction showed significant effects (**Fig S5F-G**). At the end of the *Proba* setting, the DA signals were negatively influenced by target probability, with higher probabilities resulting in lesser positive DA release for rewards and more pronounced DA decrease for omissions **(Fig 2L, Fig S5H)**. Finally, no effect of direction was observed on DA transients **(Fig S5I)**, and regarding outcome at previous trial, we observed a small effect only for rewarded trials **(Fig S5J).** Altogether, these results reveal specific patterns in the modulation of phasic DA peaks or dips across task settings. Notably, DA fluctuations were not consistently associated with the same features across rules. In *Cplx*, the current and previous outcomes explained most of the DA variations. However, the dependency on directions in the *Det* and targets in *Proba* underscores the distinct nature of DA computation in response to each of the three rules. This reinforces the idea of differences in task representation.

DA signal encodes state-specific RPEs

The observed DA fluctuations suggest a link with reward prediction errors (RPEs), which we explored through computational modeling. At each trial, we modeled DA as the sum of obtained reward (0 or 1) and RPE, adjusting RPEs trial-by-trial using the Rescorla-Wagner model **(Fig 3A, Fig S6A)**. From previous behavioral and fiber photometry results, we posited and tested three states or configurations of value representations: a simple model (M1) treating all trials equally, a model based on action (M2) with distinct values for forward and U-turn actions, and a model based on state (M3) with specific values for each target. We then used the mice's actual choices to compute model-dependent theoretical RPEs (RPE_{Mi}) and used these to fit DA variations for each mouse **(Fig 3A, Fig S6A)**. GLM analysis indicated that for each rule, only one model significantly explained DA variation, while the others two have no effect. Specifically, only M2 is significant in Det **(Fig 3B)**, only M1 in *Cplx* **(Fig 3C)**, and only M3 in Proba **(Fig 3D)**. To confirm this analysis, we show that in the Det setting only M2 was able to capture the U-turn/Forward effect observed in the fiber photometry data **(Fig 3E, Fig S6B**), and this across all learning rates tested (α, see Methods). In *Cplx*, M1 was the only model that correctly captured DA variations based on the previous outcome (**Fig 3G, Fig S6C**). Finally, in the *Proba* context, only M3, where mice learned distinct values for each target based on their probabilities, reproduced the data **(Fig 3G, Fig S6D)**. To further validate these results, we performed an extra *Proba* session, where p100 was changed into another p50. We observed that DA variations were still in line with the previous probability set, and that unexpected omissions at this new p50 target (with V_{exp} still ~1) triggered even greater DA dips **(Fig 3H)**. These findings demonstrate that mice not only learned action-value associations through DA-mediated RPE (contingency learning), but also adapted their set of relevant actions by changing their state representation from one rule to the next (representation learning).

DA dynamics adaptively reflects reward structure to foster strategy adaptation.

We next investigated how such evolution in state representation occurred within and across each rule, analyzing DA release at different phases and applying mice choice sequences to our three RL models to compute RPEs. Successive GLMs revealed evolving dominance of specific models across contexts and sessions (**Fig 4A, top**). In the Det sessions, DA variations correlated with M2 (Fwd vs Uturn) RPEs towards the end, transitioning to M1 (any trial) dominance throughout the Cplx sessions, and then progressively to M3 (p100 vs p50 vs p25) across the Proba sessions (**Fig 4A, top**). Changing the learning rate of the RL algorithm affected some statistics, without altering these patterns of evolution (**Fig S7A**). Changes in the success rates associated with each action paralleled changes in representations (**Fig 4A, bottom**), especially at transitions from one rule to another, while mice face strong discrepancies between their current internal model of the world and environmental feedbacks, requiring them to update their representation to solve a new rule. This result suggests an adaptation to changes in reward structure. Transitioning to Cplx, the success rates of all possible actions (Fwd vs Uturn, or pA vs pB vs pC) are deprecated **(Fig 4A, bottom)**, and the reward structure does not depend on specific actions but rather on the variability in the successive execution of these actions. The increase in the average success rate is actually achieved by an increase in all option-specific success rates in parallel, making a simple trial-based representation (M1) suitable to behave with this rule. When exposed to the Proba rule, mice again detect a change in the reward structure, with greater differences in success rates between locations **(Fig 4A, bottom)**, making a target-based model (M3) very efficient to represent the task, drive choice and improve performance.

To validate this interpretation, we returned to behavior to examine whether we could directly correlate concurrent evolution of decision strategy and DA dynamics. Specifically, we estimated ∆DA, the difference between DA transients associated with some options, e.g. DA(rewIpA) vs DA(rewIpB), reasoning that this ∆DA might vary with choice and performance — and thus with policy (i.e., the preference for one option among others). In *Det*, optimizing reward seeking involved reducing U-turns and sequence complexity, with no direct DA-behavior correlation **(Fig S7B-D)**. Upon transitioning to the *Cplx* rule, mice initially faced a high rate of omissions, across all available action features, due to persistence of repetitive circular choice patterns, resulting in a low success rate **(Fig S7E)**. Over time, they improved their success by increasing both U-turns and sequence complexity, generating more variability (**Fig S7E**). However, the gap in DA signals regarding previous outcome did not evolve across *Cplx*, nor did it correlate with any decision parameter **(Fig 4B)**, showing persistent differences based on reward history only **(Fig S8A-B)**. Moreover, although locally performing a Uturn led to higher chance of success (**Fig S8C**), mice did not seem to use that contingency as a heuristic: first, omissions did not locally trigger more Uturns (**Fig S8D**), and second, mice did not increase success by performing Uturns in chains, but rather by progressively learning to spread them among trials to increase variability (**Fig S8E-F**). Altogether, the results indicate that the adaptation of decision strategy in the *Cplx* rule was neither accompanied by concurrent adaptation of the DA signal nor was it a local reaction to omissions that generated negative RPEs. Upon transition to *Proba*, mice again encountered a high rate of omissions, but the distribution of those omissions was very different between possible actions, especially regarding targets **(Fig 4A, bottom)**. Across *Proba* sessions, mice progressively increased success, U-turns, and exploitation of high-probability targets **(FigS7F)**, correlating with emerging DA differences between targets **(Fig 4C)**. These concurrent adaptations, in choice preferences and in DA release, highlight independent evolution of expected values for each rewarded location. This hypothesis was confirmed by correlation analyses, demonstrating that greater divergence in DA responses to p100, p50, and p25 (higher absolute ΔDA) correlated with greater success rate, U-turns (not shown), and exploitation of high-probability targets, across both individuals and sessions **(Fig 4C).**

Discussion

By recording NAc DA release in a spatial three-armed bandit task with different rules of reward delivery, we show how DA dynamics reflected Reward Prediction Error (RPE) computations based on different task features. DA release not only conveyed value and RPE upon reward delivery or omission, but also adapted based on task contingencies, thus revealing mice internal model and representation. As the causal relationship between actions and outcomes varied across the different task rules, we hereby demonstrate that mice learned and updated values from actions (contingency learning), and changed their set of relevant states or actions from which to learn about across rules (representation learning).

First, our results confirm and extend a consistent pattern observed across the dopamine literature, wherein phasic DA carries information regarding both the obtained value and the RPE upon delivery or omission of an expected reward (*6*, *15*–*24*). More specifically, DA showed peaks in response to ICSS, regardless of whether the reward was expected or not. It remains unclear whether this response stems from direct stimulation of MFB DA fibers, resulting in DA release in the NAc, or whether it reflects a subjective value mediated by circuits beyond the DA system alone (*36*, *37*). Nevertheless, the amplitude of those peaks was modulated by task contingencies and expectations. We observed positive DA transients of greater amplitude upon unexpected rewards, and negative transients following unexpected omissions, a common observation in similar reward conditioning paradigms, interpreted as positive and negative RPEs (*6*, *7*, *24*, *38*). Using a task structured around sequential trials and choices enabled online observation of such RPE computations (both positive and negative), a phenomenon yet rarely reported (*6*, *24*, *29*, *39*, *40*), especially in the context of uncued and self-paced goal-directed decisions. These findings highlight the importance of real-time trial-based RPE measurement in detecting longitudinal changes in internal representation.

Second, mice demonstrated flexibility by switching representations and selecting relevant features to efficiently associate actions with outcomes and solve various task rules, thereby improving performance. These changes occurred during transitions between rules, when mice faced unexpected decrease in reward reward rates, suggesting that negative prediction errors and inhibition of downstream circuits by DA dips may facilitate exploration of new action representations. Under the complex rule, despite all models would have yielded similar outcomes due to the nature of the algorithm, mice opted for a specific representation that treat all trials equally, regardless of choice. The latter indicates a value-independent decision strategy, possibly together with a meta-regulation of policy parameters (for example an adaptive temperature β parameter) that promote random exploration (*8*, *41*, *42*). Upon transitioning to probabilistic setting, mice required several sessions to adjust their value representation, linking expected values to spatial preferences in a classical value-based decision-making process.

Learning rates also influenced DA variations and choice preferences. Although we used a constant rate for simplicity, learning rates might vary across contexts and individuals. Selective attention (*1*, *43*) has been proposed as an adaptive mechanism by which individuals can identify and assign credit to task-relevant features from which to learn about (*1*, *43*) possibly adjusting learning rates independently for each feature to widen the range of decision strategy adaptations. Lastly, while multiple brain areas appear to encode specific environmental features (*11*–*14*), the DA signal recorded here appeared to resolve only those features that are important for action-outcome association and used for action selection. As a result, DA dynamics could be leveraged to infer how representations are formed and how mice can flexibly adapt them to solve new rules.

References

1. Y. Niv, Learning task-state representations. Nat Neurosci 22, 1544–1553 (2019).

2. P. Dayan, Y. Niv, Reinforcement learning: The Good, The Bad and The Ugly. Current opinion in neurobiology 18, 185–196 (2008).

3. R. S. Sutton, A. G. Barto, Reinforcement Learning (MIT Press, 1998)MIT Press.

4. H. Nakahara, O. Hikosaka, Learning to represent reward structure: A key to adapting to complex environments. Neuroscience Research, 1–7 (2012).

5. B. F. Skinner, "Superstition' in the pigeon. J. Exp. Psychol. 38, 168–172 (1948).

6. T. A. Krausz, A. E. Comrie, A. E. Kahn, L. M. Frank, N. D. Daw, J. D. Berke, Dual credit assignment processes underlie dopamine signals in a complex spatial environment. Neuron, doi: 10.1016/j.neuron.2023.07.017 (2023).

7. E. Bousseyrol, S. Didienne, S. Takillah, C. Prevost-Solié, M. Come, T. A. Yahia, S. Mondoloni, E. Vicq, L. Tricoire, A. Mourot, J. Naudé, P. Faure, Dopaminergic and prefrontal dynamics co-determine mouse decisions in a spatial gambling task. Cell Rep. 42, 112523 (2023).

8. M. Belkaid, E. Bousseyrol, R. D. Cuttoli, M. Dongelmans, E. K. Duranté, T. A. Yahia, S. Didienne, B. Hanesse, M. Come, A. Mourot, J. Naudé, O. Sigaud, P. Faure, Mice adaptively generate choice variability in a deterministic task. Communications Biology 3, 1–9 (2020).

9. C. C. Beron, S. Q. Neufeld, S. W. Linderman, B. L. Sabatini, Mice exhibit stochastic and efficient action switching during probabilistic decision making. Proc National Acad Sci 119, e2113961119 (2022).

10. D. G. R. Tervo, M. Proskurin, M. Manakov, M. Kabra, A. Vollmer, K. Branson, A. Y. Karpova, Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. Cell 159, 21–32 (2014).

11. A. Banerjee, G. Parente, J. Teutsch, C. Lewis, F. F. Voigt, F. Helmchen, Value-guided remapping of sensory cortex by lateral orbitofrontal cortex. Nature 585, 245–250 (2020).

12. D. J. Barraclough, M. L. Conroy, D. Lee, Prefrontal cortex and decision making in a mixed-strategy game. Nature Neuroscience 7, 404–410 (2004).

13. S. Han, F. Helmchen, Behavior-relevant top-down cross-modal predictions in mouse neocortex. Nat. Neurosci. 27, 298–308 (2024).

14. J. Poort, A. G. Khan, M. Pachitariu, A. Nemri, I. Orsolic, J. Krupic, M. Bauza, M. Sahani, G. B. Keller, T. D. Mrsic-Flogel, S. B. Hofer, Learning Enhances Sensory and Multiple Non-sensory Representations in Primary Visual Cortex. Neuron 86, 1478–1490 (2015).

15. P. R. Montague, P. Dayan, T. J. Sejnowski, A framework for mesencephalic dopamine systems based on predictive Hebbian learning. The Journal of neuroscience : the official journal of the Society for Neuroscience 16, 1936–1947 (1996).

16. W. Schultz, P. Dayan, P. R. Montague, A Neural Substrate of Prediction and Reward. Science 275, 1593–1599 (1997).

17. N. Eshel, J. Tian, M. Bukwich, N. Uchida, Dopamine neurons share common response function for reward prediction error. Nature Neuroscience 19, 479–486 (2016).

18. H. M. Bayer, P. W. Glimcher, Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47, 129–141 (2005).

19. P. N. Tobler, C. D. Fiorillo, W. Schultz, Adaptive coding of reward value by dopamine neurons. Science (New York, N.Y.) 307, 1642–1645 (2005).

20. W. Schultz, Getting formal with dopamine and reward. Neuron 36, 241–263 (2002).

21. E. E. Steinberg, R. Keiflin, J. R. Boivin, I. B. Witten, K. Deisseroth, P. H. Janak, A causal link between prediction errors, dopamine neurons and learning. Nature Neuroscience, 1–10 (2013).

22. M. R. Roesch, D. J. Calu, G. Schoenbaum, Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nature Neuroscience 10, 1615–1624 (2007).

23. M. G. Kutlu, J. E. Zachry, P. R. Melugin, J. Tat, S. Cajigas, A. U. Isiktas, D. D. Patel, C. A. Siciliano, G. Schoenbaum, M. J. Sharpe, E. S. Calipari, Dopamine signaling in the nucleus accumbens core mediates latent inhibition. Nat. Neurosci. 25, 1071–1081 (2021).

24. A. Mohebi, J. R. Pettibone, A. A. Hamid, J.-M. T. Wong, L. T. Vinson, T. Patriarchi, L. Tian, R. T. Kennedy, J. D. Berke, Dissociable dopamine dynamics for learning and motivation. Nature 570, 65–70 (2019).

25. J. W. de Jong, Y. Liang, J. P. H. Verharen, K. M. Fraser, S. Lammel, State and rate-of-change encoding in parallel mesoaccumbal dopamine pathways. Nat. Neurosci. 27, 309–318 (2024).

26. S. J. Gershman, J. A. Assad, S. R. Datta, S. W. Linderman, B. L. Sabatini, N. Uchida, L. Wilbrecht, Explaining dopamine through prediction errors and beyond. Nat. Neurosci., 1–11 (2024).

27. A. Lak, W. R. Stauffer, W. Schultz, Dopamine prediction error responses integrate subjective value from different reward dimensions. Proceedings of the National Academy of Sciences of the United States of America 111, 2343–2348 (2014).

28. C. D. Fiorillo, P. N. Tobler, W. Schultz, Discrete coding of reward probability and uncertainty by dopamine neurons. Science (New York, N.Y.) 299, 1898–1902 (2003).

29. A. A. Hamid, J. R. Pettibone, O. S. Mabrouk, V. L. Hetrick, R. Schmidt, C. M. V. Weele, R. T. Kennedy, B. J. Aragona, J. D. Berke, Mesolimbic dopamine signals the value of work. Nature Neuroscience, doi: 10.1038/nn.4173 (2015).

30. E. S. Bromberg-Martin, M. Matsumoto, O. Hikosaka, Dopamine in motivational control: rewarding, aversive, and alerting. Neuron 68, 815–834 (2010).

31. E. S. Bromberg-Martin, O. Hikosaka, Midbrain dopamine neurons signal preference for advance information about upcoming rewards. Neuron 63, 119–126 (2009).

32. L. T. Coddington, J. T. Dudman, The timing of action determines reward prediction signals in identified midbrain dopamine neurons. Nat Neurosci 21, 1563–1573 (2018).

33. J. D. Berke, What does dopamine mean? Nat Neurosci 21, 787–793 (2018).

34. M. Dongelmans, R. D. Cuttoli, C. Nguyen, M. Come, E. K. Duranté, D. Lemoine, R. Brito, T. A. Yahia, S. Mondoloni, S. Didienne, E. Bousseyrol, B. Hannesse, L. M. Reynolds, N. Torquet, D. Dalkara, F. Marti, A. Mourot, J. Naudé, P. Faure, Chronic nicotine increases midbrain dopamine neuron activity and biases individual strategies towards reduced exploration in mice. Nat Commun 12, 6945 (2021).

35. J. Naudé, S. Tolu, M. Dongelmans, N. Torquet, S. Valverde, G. Rodriguez, S. Pons, U. Maskos, A. Mourot, F. Marti, P. Faure, Nicotinic receptors in the ventral tegmental area promote uncertainty-seeking. Nature Neuroscience 19, 471–478 (2016).

36. I. Trujillo-Pisanty, K. Conover, P. Solis, D. Palacios, P. Shizgal, Dopamine neurons do not constitute an obligatory stage in the final common path for the evaluation and pursuit of brain stimulation reward. PLoS ONE 15, e0226722 (2020).

37. S. J. Millard, I. B. Hoang, S. Sherwood, M. Taira, V. Reyes, Z. Greer, S. L. O'Connor, K. M. Wassum, M. H. James, D. J. Barker, M. J. Sharpe, Cognitive representations of intracranial self-

stimulation of midbrain dopamine neurons depend on stimulation frequency. Nat. Neurosci. 27, 1253–1259 (2024).

38. M. Blanco-Pozo, T. Akam, M. E. Walton, Dopamine-independent effect of rewards on choices through hidden-state inference. Nat. Neurosci., 1–12 (2024).

39. L. T. Coddington, S. E. Lindo, J. T. Dudman, Mesolimbic dopamine adapts the rate of learning from action. Nature, 1–9 (2023).

40. M. G. Kutlu, J. E. Zachry, P. R. Melugin, S. A. Cajigas, M. F. Chevee, S. J. Kelly, B. Kutlu, L. Tian, C. A. Siciliano, E. S. Calipari, Dopamine release in the nucleus accumbens core signals perceived saliency. Curr. Biol. 31, 4748-4761.e8 (2021).

41. M. Dubois, J. Habicht, J. Michely, R. Moran, R. J. Dolan, T. U. Hauser, Human complex exploration strategies are enriched by noradrenaline-modulated heuristics. eLife 10 (2021).

42. R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, J. D. Cohen, Humans use directed and random exploration to solve the explore-exploit dilemma. Journal of experimental psychology. General 143, 2074–2081 (2014).

43. S. J. Gershman, K. A. Norman, Y. Niv, Discovering latent causes in reinforcement learning. Curr. Opin. Behav. Sci. 5, 43–50 (2015).

Acknowledgements:

This work was supported by the Centre National de la Recherche Scientifique CNRS UMR 8246 and 8249, INSERM U1130. We are grateful to the animal facilities (IBPS) and Otilia de Oliveira and Emilie Tubeuf at ESPCI animal facilities. We also thank Jérémie Naudé and Clément Solié for their comments on the manuscript.

Fundings:

The Foundation for Medical Research (FRM, Equipe FRM DEQ2013326488 to P.F), the French National Cancer Institute Grant TABAC-16-01, TABAC-19-020 and SPA-21-002 (to P.F.). French state funds managed by the ANR (ANR-19-CE16-0028 Bavar to PF). Fourth-year PhD fellowship from the Foundation for Medical Research (FDT201904008060 to MC).

Authors contributions:

Conceptualization: MC, PF

Injection and implantation surgeries: MC

Behavioral experiments: MC, AG, LK

Fiber photometry recordings: MC, AG

Intracardiac perfusions and immunohistochemistry: MC, AG, EV, TLB

Data analysis: MC, PF

Modelling: MC, PF

Setups development: MC, JJ, AM, EB, SD, PF

Funding acquisition: PF

Writing - original draft: MC, PF

Writing - review and editing: MC, TLB, AM, PF

Competing interests:

Authors declare that they have no competing interests

Data and materials availability: All the data that support the findings of this study can be found in the Source Data file provided with the paper. If necessary, the raw data from the online behavioral experiment (i.e the trajectories) are available from the corresponding author. All codes used to run the analysis are available from the authors upon request.

Supplementary Materials

Materials and Methods

Figs. S1 to S8

Tables of detailed statistics for Figs. 1-4 and Supp S1-S8

References

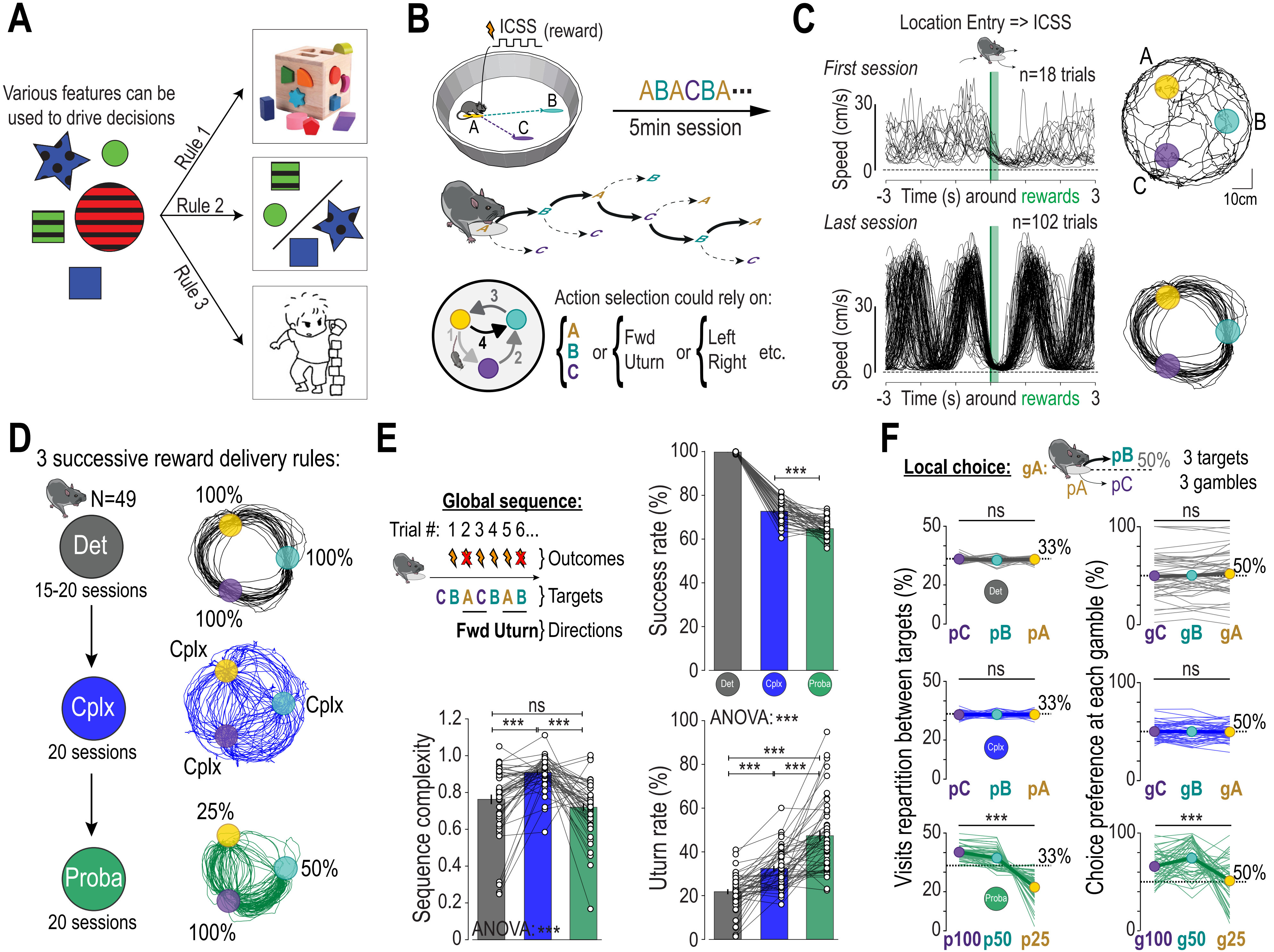

A
Various features can be used to drive decisions
Rule 1
Rule 2
Rule 3
B
ICSS (reward)
A
B
C
ABACBA···
5min session
3
1
4
2
Action selection could rely on:
A B C
or
Fwd Uturn
or
Left Right
etc.
C
Location Entry => ICSS
First session
n=18 trials
Speed (cm/s)
30
0
-3 Time (s) around rewards 3
Last session
n=102 trials
A
B
C
10cm
D
3 successive reward delivery rules:
N=49
Det
15-20 sessions
Cplx
20 sessions
Proba
20 sessions
100%
100%
100%
Cplx
Cplx
Cplx
25%
50%
100%
E
Global sequence:
Trial #: 1 2 3 4 5 6...
Outcomes
C B A C B A B
Targets
Fwd Uturn
Directions
Success rate (%)
100
80
60
40
20
0
***
Det
Cplx
Proba
Sequence complexity
1.2
1.0
0.8
0.6
0.4
0.2
0
ns
***
***
ANOVA: ***
Uturn rate (%)
ANOVA: ***
***
***
***
F
Local choice:
gA:
pB 50%
pA
pC
3 targets
3 gambles
Visits repartition between targets (%)
Choice preference at each gamble (%)
50
20
0
ns
33%
pC pB pA
100
50%
gC gB gA
***
p100 p50 p25
g100 g50 g25

Figure 1

Fig. 1. Mice display distinct reward seeking strategies adapted to each rule. A. From a variety of overlapping features, individuals can learn value and take decisions depending on the rule. **B.** Mice perform successive binary choices to collect ICSS rewards. Choice could rely on various overlapping sets of actions. **C.** Speed profiles and trajectories throughout conditioning. **D.** Three reward delivery rules were successively proposed: Deterministic (*Det*) where all trials were rewarded (P=100%), Complexity (*Cplx*), where trials are rewarded based on sequence variability, and Probabilistic (*Proba*), with each target associated to a given probability (P=25%, 50%, and 100%). **E.** Succession of trials and choices generates sequences of outcomes (rewards and omissions), targets (A, B and C) and directions (Forwards and Uturns). Comparison of success rate, sequence complexity and Uturn rate reveals distinct reward seeking strategies across contexts. **F.** Locally, a mouse on one location (ex: p_A) has the choice between the two others (ex: p_B vs p_C), and therefore performs a gamble computed as $g_A = P(p_B|p_A)$. g>50% corresponds to clockwise rotation for *Det* and *Cplx*, and to preference for highest probability of reward for *Proba*. Proportion of target visits and choice preference at each gamble show a bias for circular foraging in Det, exploitation in *Proba*, and randomness in *Cplx*. Data are shown as individual points, and mean ±sem. N=49 mice (23 males and 26 females).

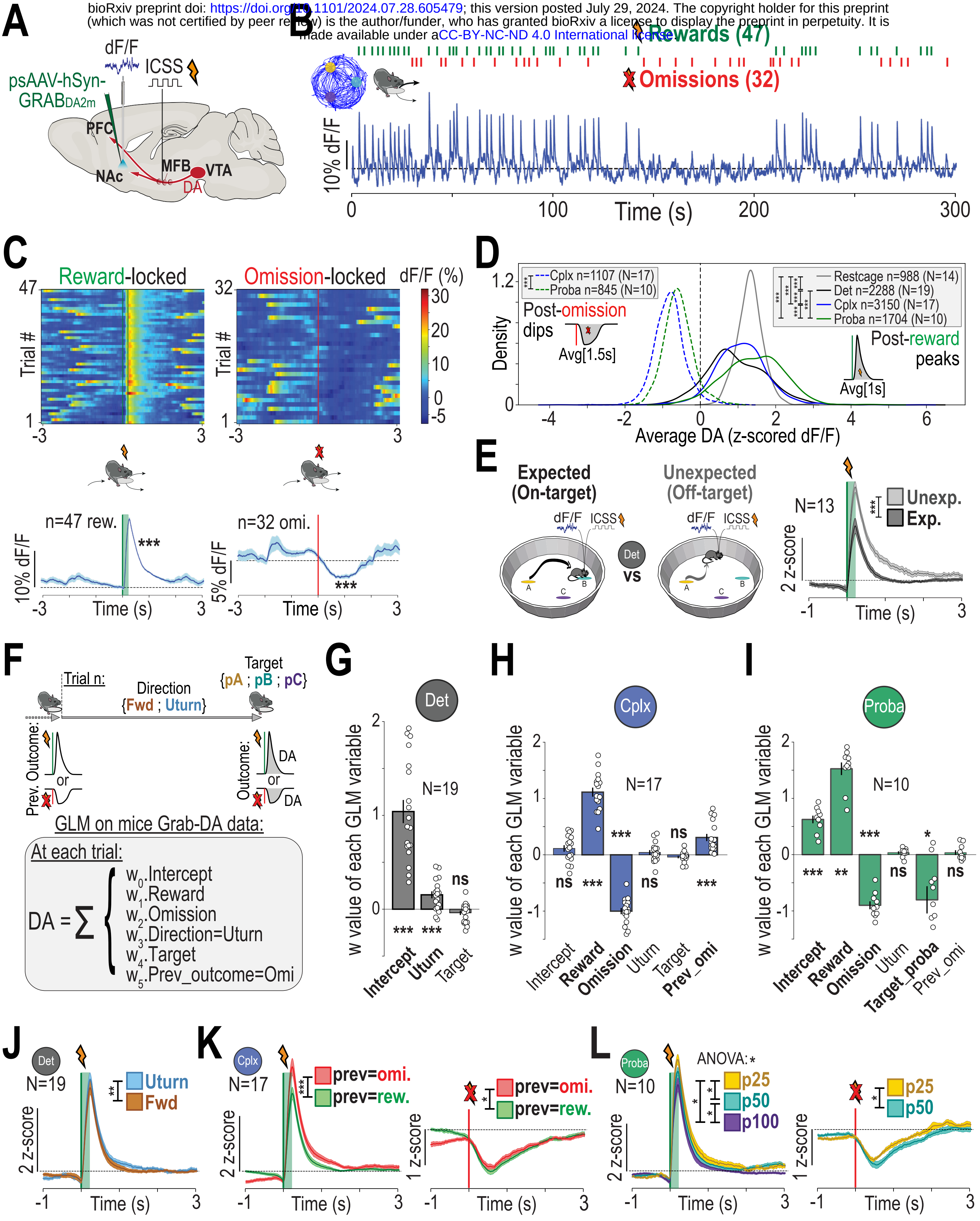

A
dF/F
ICSS
psAAV-hSyn-GRAB$_{DA2m}$
PFC
NAc
MFB
VTA
DA
B
Rewards (47)
Omissions (32)
10% dF/F
0
100
Time (s)
200
300
C
Reward-locked
Omission-locked
dF/F (%)
47
32
Trial #
1
-3
3
30
20
10
0
-5
n=47 rew.
***
n=32 omi.
***
10% dF/F
5% dF/F
Time (s)
D
Cplx n=1107 (N=17)
Proba n=845 (N=10)
***
Post-omission dips
Avg[1.5s]
Restcage n=988 (N=14)
Det n=2288 (N=19)
Cplx n=3150 (N=17)
Proba n=1704 (N=10)
Post-reward peaks
Avg[1s]
1.2
Density
0
-4
-2
0
2
4
6
Average DA (z-scored dF/F)
E
Expected (On-target)
Unexpected (Off-target)
dF/F
ICSS
A
B
C
Det
vs
N=13
***
Unexp.
Exp.
2 z-score
-1
Time (s)
3
F
Trial n:
Target
{pA ; pB ; pC}
Direction
{Fwd ; Uturn}
Prev. Outcome:
or
Outcome:
DA
GLM on mice Grab-DA data:
At each trial:
DA = ∑ { w$_0$.Intercept, w$_1$.Reward, w$_2$.Omission, w$_3$.Direction=Uturn, w$_4$.Target, w$_5$.Prev_outcome=Omi }
G
Det
N=19
w value of each GLM variable
2
1
0
*** *** ns
Intercept
Uturn
Target
H
Cplx
N=17
2
1
0
-1
ns *** *** ns ns ***
Intercept
Reward
Omission
Uturn
Target
Prev_omi
I
Proba
N=10
2
1
0
-1
*** ** *** ns * ns
Intercept
Reward
Omission
Uturn
Target_proba
Prev_omi
J
Det
N=19
**
Uturn
Fwd
2 z-score
-1
Time (s)
3
K
Cplx
N=17
***
prev=omi.
prev=rew.
*
prev=omi.
prev=rew.
2 z-score
1 z-score
-1
Time (s)
3
L
Proba
N=10
ANOVA:*
*
*
*
p25
p50
p100
*
p25
p50
2 z-score
1 z-score
-1
Time (s)
3

Figure 2

Fig. 2: NAc DA release dynamics reveal expectations built upon rule-specific features. A. Schematic of the experimental design to record DA release during the task with chronic fiber photometry. **B.** Representative signal from one 5-min session. **C.** For the same example session, signal is time-locked on location entry (t0) and averaged. Rewards induce peaks and omissions induce dips of DA release. **D.** Density distribution of averaged DA variations for rewards and omissions for the last two sessions of Det, Cplx or Proba, and for random stimulations in the rest cage (performed on last day of Det). **E.** After conditioning, mice were randomly and unexpectedly stimulated during the task outside of the rewarded zones (off-target), triggering DA peaks of greater amplitude. **F.** Each trial is defined by predictors (outcome received, previous outcome received, trajectory chosen to reach target, and target chosen) to fit DA amplitude using GLMs. **G-H-I.** GLM results at the end of Det, Cplx and Proba. Features explaining DA variations vary across contexts. **J-K-L.** Direct analysis of DA transients locked on those significant features (Uturn vs Fwd in Det ; reward vs omission at previous trial in Cplx ; p25 vs p50 vs p100 in Proba). Data are shown as individual points, and/or mean ±sem. n is the number of trials, N the number of mice in each condition.

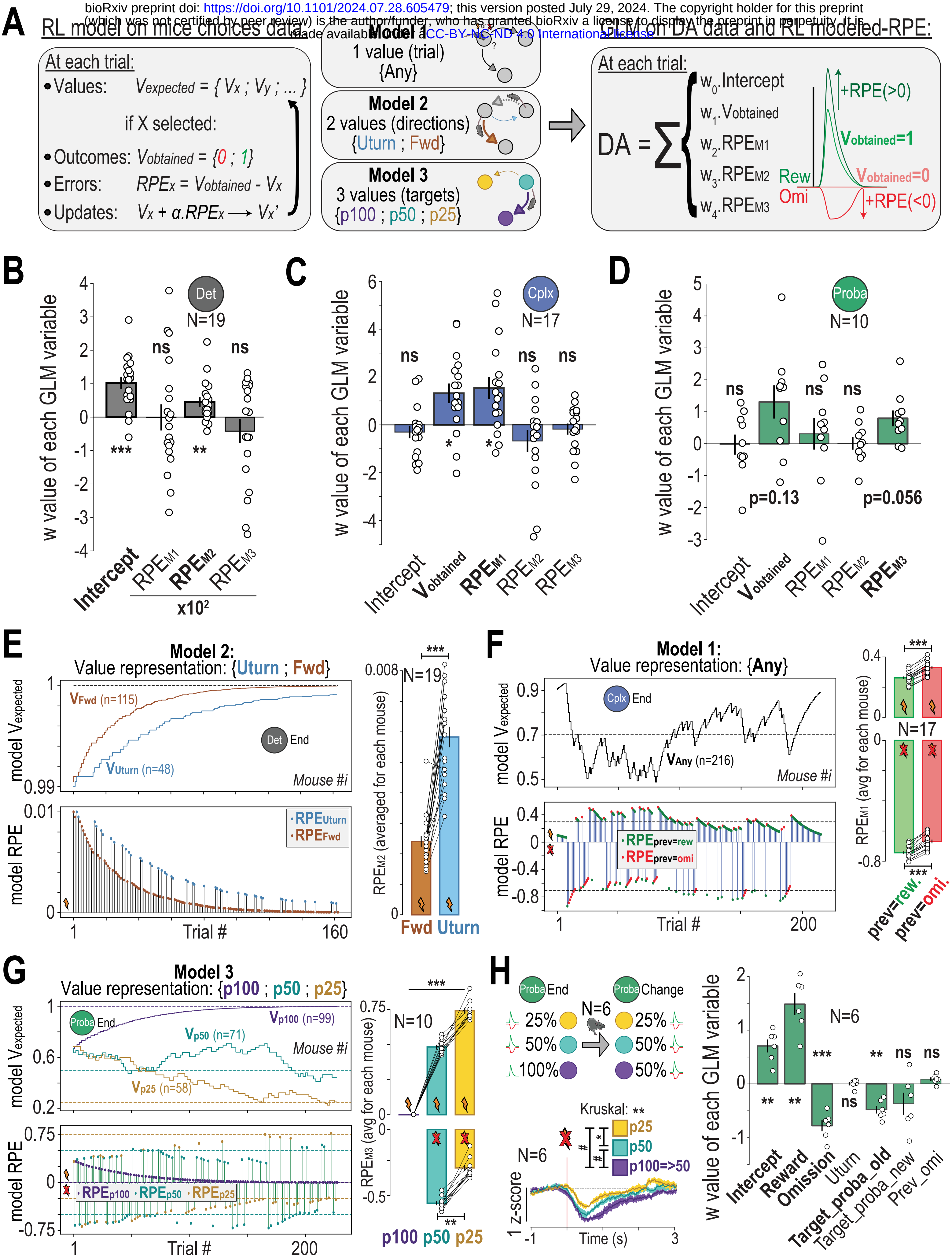

A
RL model on mice choices data:
At each trial:
• Values: Vexpected = { Vx ; Vy ; ... }
if X selected:
• Outcomes: Vobtained = {0 ; 1}
• Errors: RPEx = Vobtained - Vx
• Updates: Vx + α.RPEx ⟶ Vx'
Model 1
1 value (trial)
{Any}
Model 2
2 values (directions)
{Uturn ; Fwd}
Model 3
3 values (targets)
{p100 ; p50 ; p25}
GLM on DA data and RL modeled-RPE:
At each trial:
DA = ∑
w0.Intercept
w1.Vobtained
w2.RPEM1
w3.RPEM2
w4.RPEM3
+RPE(>0)
Vobtained=1
Rew
Omi
Vobtained=0
+RPE(<0)
B
w value of each GLM variable
Det
N=19
***
ns
**
ns
Intercept
RPEM1
RPEM2
RPEM3
x10²
C
Cplx
N=17
ns
*
*
ns
ns
Intercept
Vobtained
RPEM1
RPEM2
RPEM3
D
Proba
N=10
ns
ns
ns
p=0.13
p=0.056
Intercept
Vobtained
RPEM1
RPEM2
RPEM3
E
Model 2:
Value representation: {Uturn ; Fwd}
model Vexpected
VFwd (n=115)
VUturn (n=48)
Det End
Mouse #i
model RPE
RPEUturn
RPEFwd
Trial #
1
160
RPEM2 (averaged for each mouse)
0.008
N=19
***
Fwd
Uturn
F
Model 1:
Value representation: {Any}
model Vexpected
Cplx End
VAny (n=216)
Mouse #i
model RPE
RPEprev=rew
RPEprev=omi
Trial #
1
200
RPEM1 (avg for each mouse)
***
N=17
***
prev=rew.
prev=omi.
G
Model 3
Value representation: {p100 ; p50 ; p25}
model Vexpected
Proba End
Vp100 (n=99)
Vp50 (n=71)
Vp25 (n=58)
Mouse #i
model RPE
RPEp100
RPEp50
RPEp25
Trial #
1
200
RPEM3 (avg for each mouse)
***
N=10
**
p100
p50
p25
H
Proba End
Proba Change
N=6
25%
50%
100%
25%
50%
50%
Kruskal: **
p25
p50
p100=>50
N=6
1 z-score
Time (s)
-1
3
w value of each GLM variable
N=6
**
**
***
ns
**
ns
ns
Intercept
Reward
Omission
Uturn
Target_proba_old
Target_proba_new
Prev_omi

Figure 3

Fig. 3: DA signal embeds an RPE component, modelled from distinct value representations specific to each rule. A. Mice choice sequences were taken to train Reinforcement Learning (RL) algorithms, testing three possible action representations to update values and compute corresponding RPEs. Model 1 (M1) treats all trials equally with fluctuating { V_{Any} }. M2 updates a set of two distinct values { V_{Fwd} ; V_{Uturn} }. A spatial model (M3) computes three independent values for each target { V_{pA} ; V_{pB} ; V_{pC} }. We then trained another GLM assuming DA = $V_{obtained}$ + RPE, with trial RPEs generated from M1, M2 and M3. **B-C-D.** GLM results in Det, Cplx and Proba. Models reproducing RPEs that explained DA variations vary across contexts. **E-F-G.** Evolution of expected value and RPE for M1, M2 or M3 in example sessions (left) and on average (right). **E.** In M2-Det, convergence toward 1 is slower for Uturns, leading to higher RPE_{Uturn} and reproducing DA data. **F.** In M1-Cplx, V_{Any} is always updated and fluctuates around mean success rate. Plotting corresponding RPEs regarding current and previous outcomes mimic DA data. **G.** In M3-Proba, value of each target converges and then fluctuates around its probability, and corresponding RPEs reproduce DA data. **H.** At the end of Proba, probability of the p100 location was changed to 50%. Omissions at target p100=>50 triggered deeper DA dips, while GLM shows DA still varies with the old probability set. Data are shown as individual points, and/or mean ±sem. n is the number of trials, N the number of mice in each condition.

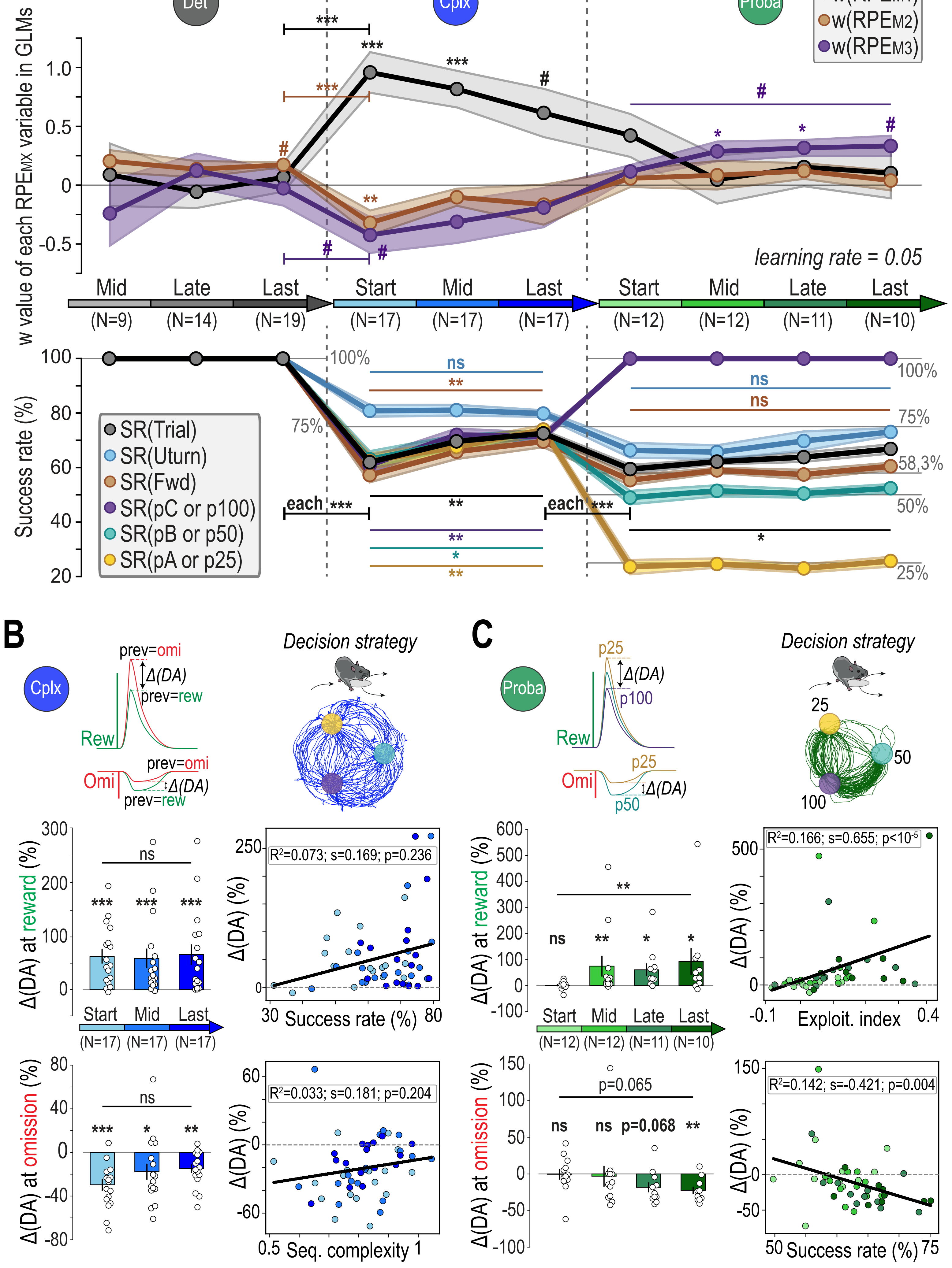

A
Det
Cplx
Proba
w(RPEM1)
w(RPEM2)
w(RPEM3)
w value of each RPEMX variable in GLMs
1.0
0.5
0
-0.5
***
***
***
#
#
*
*
#
#
#
**
#
learning rate = 0.05
Mid
Late
Last
Start
Mid
Last
Start
Mid
Late
Last
(N=9)
(N=14)
(N=19)
(N=17)
(N=17)
(N=17)
(N=12)
(N=12)
(N=11)
(N=10)
Success rate (%)
100
80
60
40
20
100%
75%
58,3%
50%
25%
ns
**
ns
ns
SR(Trial)
SR(Uturn)
SR(Fwd)
SR(pC or p100)
SR(pB or p50)
SR(pA or p25)
each ***
**
each ***
**
*
**
*
B
Cplx
prev=omi
Δ(DA)
prev=rew
Rew
Omi
prev=omi
Δ(DA)
prev=rew
Decision strategy
Δ(DA) at reward (%)
300
200
100
0
ns
***
***
***
Start
Mid
Last
(N=17)
(N=17)
(N=17)
Δ(DA) (%)
250
0
R²=0.073; s=0.169; p=0.236
30
Success rate (%)
80
Δ(DA) at omission (%)
80
40
0
-40
-80
ns
***
*
**
Δ(DA) (%)
60
0
-60
R²=0.033; s=0.181; p=0.204
0.5
Seq. complexity
1
C
Proba
p25
Δ(DA)
p100
Rew
Omi
p25
Δ(DA)
p50
Decision strategy
25
50
100
Δ(DA) at reward (%)
600
500
400
300
200
100
0
-100
**
ns
**
*
*
Start
Mid
Late
Last
(N=12)
(N=12)
(N=11)
(N=10)
Δ(DA) (%)
500
0
R²=0.166; s=0.655; p<10⁻⁵
-0.1
Exploit. index
0.4
Δ(DA) at omission (%)
150
100
50
0
-50
-100
p=0.065
ns
ns
p=0.068
**
Δ(DA) (%)
150
0
-50
R²=0.142; s=-0.421; p=0.004
50
Success rate (%)
75

Figure 4

Fig. 4: DA reflects switches in task representations, fostering strategy adaptation to improve performance. A. Top: The approach of RL modelling and GLM fitting DA data with computed RPEs used in Fig 3 was extended at different phases across each rule. To mimic mice learning, we took the final values of models at phase n to feed the initial values of models at phase n+1. Plots show evolution of RPE_{M1} (black), RPE_{M2} (brown) and RPE_{M3} (purple) weights over time. Bottom: Parallel general or action-related success rates. Rule transitions represent high degrees of discrepancy. **B.** ΔDA is computed for each session of each mouse as the relative difference Δ = (prev_omi—prev_rew) / prev_rew, for both rewards and omissions, showing significant effect of previous outcome for all phases in *Cplx*, but with neither ΔDA adaptation across sessions, nor correlation with any decision parameter across sessions and individuals. **C.** Same as B, but computing ΔDA as difference between high and low probability targets in *Proba*. ΔDA adapts throughout the *Proba* sessions, with strong correlations with decision parameters. (Data are shown as individual points, and/or mean ± sem. In B, C linear regressions, each data point is one animal at one phase. In A, due to multiple corrections (x10) generating dilutions in p-values, **#** symbol has been added to highlight $p<0.12$ after correction. N is always the number of mice in each context.

Supplementary Materials

Materials and Methods

Animals: Experiments were performed on adult C57Bl/6Rj wild-type mice (Janvier Labs, France). Both male and female mice, weighing 20-30 g and 8 weeks old at the time of surgery, were used for behavioral experiments. Only male mice were used in the $GRAB_{DA}$ fiber photometry cohorts. For cre-dependent GCaMP experiments, DATiCre male mice were used. All mice were kept in an animal facility where temperature (20 ± 2°C) and humidity were automatically monitored and a circadian 12/12h light–dark cycle was maintained. All experiments were performed in accordance with the recommendations for animal experiments issued by the European Commission directives 219/1990, 220/1990, and 2010/63, and approved by Sorbonne University and ESPCI.

AAV production: AAVs for $GRAB_{DA2m}$ (pXR1-AAV-hSyn-GRAB-DA4.4) were produced as previously described (*1*) (using the cotransfection method from plasmids generously provided by Dr. Yulong Lee (*2*, *3*) and purified by iodixanol gradient ultracentrifugation(*4*)). AAV vector stocks were tittered by quantitative PCR (qPCR) (*5*) using SYBR Green (Thermo Fischer Scientific). AAV vectors for GCaMP6f (AAV1-EF1a-DIO-GCaMP6f-P2A-nls-dTomato) and GCaMP7c (pGP-AAV1-syn-FLEX-jGCaMP7c variant 1513-WPRE) were directly ordered from Addgene.

Intracranial self-stimulation (ICSS) electrode implantation: Male and female WT mice were anaesthetized with a gas mixture of oxygen (1 L/min) and 1-3% of isoflurane (Piramal Healthcare, UK) and then placed into a stereotaxic frame (Kopf Instruments, CA, USA). After the administration of a local anesthetic (Lurocain, 0.1 mL at 0.67 mg/kg), a median incision revealed the skull, which was drilled at the level of the median forebrain bundle (MFB). For ICSS, a bipolar stimulating electrode (PlasticOne 2 channels, stainless steel, 10 mm) was then implanted unilaterally (left or right, randomized) in the brain using the following stereotaxic coordinates (from bregma according to Paxinos atlas): AP −1.4 mm, ML ±1.2 mm, DV −4.8 mm from the brain). Dental cement (SuperBond, Sun Medical) was used to fix the implant to the skull. An analgesic solution of buprenorphine at 0.015 mg/L (0.1 mL/10 g) was delivered prior to awakening from the surgery and, if necessary, the following recovering days. After stitching, mice were placed back in their home-cage and had a minimum of 5 days to recover from surgery. The efficacy of electrical stimulation was verified through the rate of conditioning during the deterministic setting (see Intracranial Self Stimulation (ICSS) bandit task). Out of the 54 mice implanted (27 for each sex), 49 were included in the results (23 males and 26 females).

Virus injections and fiber photometry recordings: 3 cohorts of WT male mice (total of 24) were anaesthetized (Oxygen 1 L/min, Isoflurane 1–3%) and implanted with an ICSS electrode as described above. They were then injected unilaterally (randomized left/right side and ipsi/contralateral side regarding the ICSS electrode) in the NAc lateral shell (1 µL, coordinates from bregma: AP +1.45mm; ML ±1.55mm; DV −4.05mm from the skull) with an adeno-associated virus (*2*, *3*) to express GRAB$_{\mathrm{DA2m}}$. An optical fiber (200 µm core, NA = 0.39, Thor Labs) coupled to a metallic ferule (1.25 mm) was implanted 100 µm above the injection site in target region and cemented to the skull with blackened cement. 5 DATiCre male mice followed the same procedures for GCaMP experiments in the VTA (1 µL, coordinates from bregma: AP -3.10mm; ML ±0.50mm; DV −4.20mm from the brain), 3 of them with GCaMP7c and 2 with GCaMP6f. Viral expression typically took 10-15 days to achieve a satisfying signal and lasted for up to 3 months. However, some mice exhibited a shorter duration of expression and were therefore excluded for the analysis of later sessions. Although the mice performed the task on a daily basis, fluorescent recordings were made only every 2 or 3 days to prevent sensor bleaching. Low power (100-200 mA) LEDs (465 nm and 405 nm, Doric Lenses) coupled to a patch cord (500 µm core, NA = 0.5, Prizmatix) were used for optical stimulation of the sensors in lock-in mode (572.205 Hz for the 465 nm LED, 208.616 Hz for the 405 nm LED) and collection of 520 nm fluorescence. 405 nm was used as the isobestic wavelength. The optical stimulation patch cord was plugged onto the ferrule during all experimental sessions, even those without recordings, to habituate animals and control for latent experimental effects. After the daily session, a short recording of the autofluorescence signal $F(auto)$, coming from the patchcord only, was performed with same LED intensities, no animal plugged and room in the dark. Raw 520 nm fluorescence was demodulated by the software (Doric Lenses) to extract 465 nm and 405 nm signals. The 405 nm signal was visually checked to account for instability artefacts coming from head movements or patch cord unplugging during the session, and if needed correct the associated 465 nm signal accordingly, otherwise it was not used for signal treatment. 465 nm signal F_i follows several treatment steps according to this formula:

$$\frac{dF_i}{F_0} = \frac{F_i - F(auto)\ -\ F_i(fit)}{F_i(fit)}\ -\ 1$$

First F_i is subtracted with the constant value of autofluorescence *F(auto)* measured with patch cord only, improving drastically the signal-to-noise ratio. Then, largest transients induced by ICSS were excluded to perform a smoothing on the subsequent truncated signal. We then computed a mono-exponential fit $F_i(fit)$ on this smoothed signal, which was also subtracted to F_i at each time point i to account for exponential decay. The result is then divided by the same $F_i(fit)$ at each time point i to normalize the signal around 1, and subtracted by the constant 1 to normalize to 0 and obtain positive or negative transients as dF_i/F_0 over an entire session (5 or 10min). In order to aggregate signals coming from different sessions for each mouse, and then pool mice for the analysis, we also applied a z-scoring on dF_i/F_0 over each entire session.

Intracranial self-stimulation (ICSS) bandit task The ICSS bandit task (*6*–*9*), took place in a circular open-field with a diameter of 68 cm. Three explicit square-shaped marks (2 × 2 cm) were taped in the open field, forming an equilateral triangle (side = 35 cm). Entry in the circular zones (diameter = 6 cm) around each mark was associated with the delivery of a rewarding ICSS stimulation. A LabVIEW (National Instruments) application precisely tracked and recorded the animal's position with a camera (20 frames/s). When a mouse was detected in one of the circular rewarding zones, a TTL signal was sent to the electrical stimulator, which generated a 200 ms train of 5 ms biphasic square waves pulsed at 100 Hz (20 pulses per train). Two consecutive rewards could not be delivered on the same target, which motivated mice to alternate between targets and therefore generate sequences of binary choices. ICSS intensity was adjusted, within a range of 15-200 μA, during early conditioning sessions, so that mice would achieve between 50 and 120 visits per session (5 min duration) for two successive sessions. ICSS intensity was then kept constant for all the experiments, even when reward delivery rules changed. Mice with insufficient scores were excluded. Different reward delivery rules were used, and all animals went through all three protocols successively. The first is a deterministic (Det) setting, with 10 to 15 daily sessions of 5 min. All zones were associated with an ICSS delivery (P = 100%). The second, described previously in (*6*), is a complex (Cplx) setting where a grammatical complexity algorithm (*10*) analyses online the choice sequence that the mouse is producing, calculates the complexity of two potential sequences of length 10 (9 past targets + next target among the 2 available) and gives a reward only if the complexity of the sequence increases. Repeating patterns of low complexity will therefore lead to series of omissions, while increasing variability will increase success rate. Mice did daily sessions during 15-20 days. The third setting is probabilistic (Proba): each target is associated with a probability to obtain an ICSS stimulation among three (P = 25%, P = 50%, P = 100%), as described previously (*7*–*9*). The probabilities at each location were pseudo-randomly assigned per mouse, and 15-20 sessions were performed. 2 cohorts of both male and female mice followed deterministic, complexity and probability settings successively, with no fluorescent sensor expression. Three cohorts of male mice expressing $GRAB_{DA}$ and implanted with an optical fiber implantation followed different settings: *i)* the first cohort performed only Det and Cplx, and recordings started only at the end of Det, *ii)* the second and third cohorts performed Det, Cplx and Proba, with recordings starting at the beginning of Det, and performed also some control experiments (especially, unexpected rest cage and off-target ICSS). Consequently, there is variation in animal numbers among conditions in the figures. Finally, one cohort of DATiCre male mice was tested in Det and Cplx only.

Behavioral measures: For all those groups, the following measures were analyzed with custom codes in Python (using mostly Numpy and Pandas libraries, on PyCharm CE) and compared throughout the different rules: *i)* number of visits, *ii)* success rate, *iii)* time-to-goal, *iv)* choice repartition (proportion of visits at each location), *v)* percentage of U-turn (target n = target n+2) and *vi)* sequence complexity (applying the same complexity algorithm calculation but offline and for all choices during a session). Furthermore, the ICSS bandit task can be seen as a Markovian decision process: every transition can be considered as a binary choice between two options, since a zone cannot be reinforced twice in a row. The sequence of choices per session results from the succession of three specific binary choices, or gambles. For deterministic and complexity, G_C = P(A|C) would be the total number of visits in target A divided by the total number of visits in targets A and B, when the animal is in target C. Similarly, G_A = P(B|A) and G_B = P(C|B). A gamble above 50% indicates that the animal has a preference for moving clockwise (or below 50% for moving counter-clockwise). In probabilistic, direction of conditional probabilities does not follow spatial repartition of locations, but rather preference for the high value option: G_{25} = 100% vs 50%, G_{100} = 50% vs 25% and G_{50} = 100% vs 25%. Applying this principle at each choice, those 3 gambles can be aggregated into single values to give circularity index (going in circle, no matter clockwise or counter-clockwise), exploitation index (always preferring the highest value option) or repetition index (always making the same choice at given gamble, no matter the direction or exploitation).

Fiber photometry analysis: All treatments and analyses were performed in Python using custom codes (mostly Numpy and Pandas libraries). After cleaning and processing each session signal to obtain dF/F values and z-scored dF/F values, events of interest were extracted to align the signal in [-3s:3s] time window in dataframes, t_0 being the exact time of location entry (triggering reward delivery or omission), with 1kHz sampling. Session-wise averages of given conditions for each mouse were then extracted, and averaged again over multiple mice for statistical analyses. In some conditions, especially when events of interest were rare (some scenarios of rewards or omissions chains in complexity, or some scenarios of locations transition in probabilistic), two or more sessions from one animal were pooled as if they were one (for instance, the last two sessions in a given context) to have enough trials for each animal in this condition. For the same reason, the third cohort of $GRAB_{DA}$ mice followed 10 min long sessions (instead of 5 min) in Cplx and Proba settings, with no particular effect on the overall quality of the signal, nor the duration of $GRAB_{DA}$ expression (up to 3 months). For $GRAB_{DA}$, rewards-elicited positive transients typically peaked around 250 ms after location entry (duration of ICSS being 200 ms) and decayed during a bit less than 1s: we therefore extracted maximum and mean of the signal in a 1 s window post location entry. Omissions-elicited negative transients were longer, reaching their minimum around 800 ms after location entry and taking roughly 700-800 ms to go back to baseline: we therefore extracted minimum and mean of the signal in 1.5 s window post location entry. For GCaMP, kinetics depended on the sensor used: peaks reached maximum value around 250-300 ms post location entry for GCaMP6f and 350-400 ms for GCaMP7c, while dips reached minimum value around the same time (900-1050 ms post location entry) for both sensors. However, return to baseline after reward-induced peaks was much shorter for GCaMP6f (500-600 ms post location entry) than for GCaMP7c (2-3 s). For some correlation analyses (using SciKit Learn Python library), especially the ones regarding z-scored peaks or dips amplitude regarding outcome chain history, all trials of all mice were pooled together in a given condition.

Generalised Linear Model (GLM) approach: GLM was performed in Python using custom codes (StatsModels or SciKit Learn library). To disentangle multiple factors that could explain DA signal, due to high degree of behavioral and task-related variables correlated to each other from one trial to the next, we designed a generalized linear model where a variable Y is explained by a linear combination of multiples variables X_i, each of them weighted by a parameter w_i, plus a residual (or intercept) w_0.

$$Y = \mathrm{w}_0 + \mathrm{w}_1 . X_1 + \mathrm{w}_2 . X_2 + \ldots$$

The model aims at fitting variations of Y by determining the weights w_i and their significance. Dependent variable Y was post location entry 1s average for reward-induced peaks or 1.5s average for omission-induced dips. Multiple X_i variables have been used, namely: *i)* reward or omission at previous and current location, *ii)* Forward or U-turn at previous trial, *iii)* current target visited (spatially A, B or C, or in Proba p_{100}, p_{50} or p_{25}), and *iv)* time since last stimulation (in Restcage stimulation condition). A single GLM was applied for each mouse in a given condition, then w_i parameters resulting from all those GLMs were averaged among mice, and the average was statistically compared to 0. Significance, either with positive or negative weight, indicates that this variable explains part of DA variations.

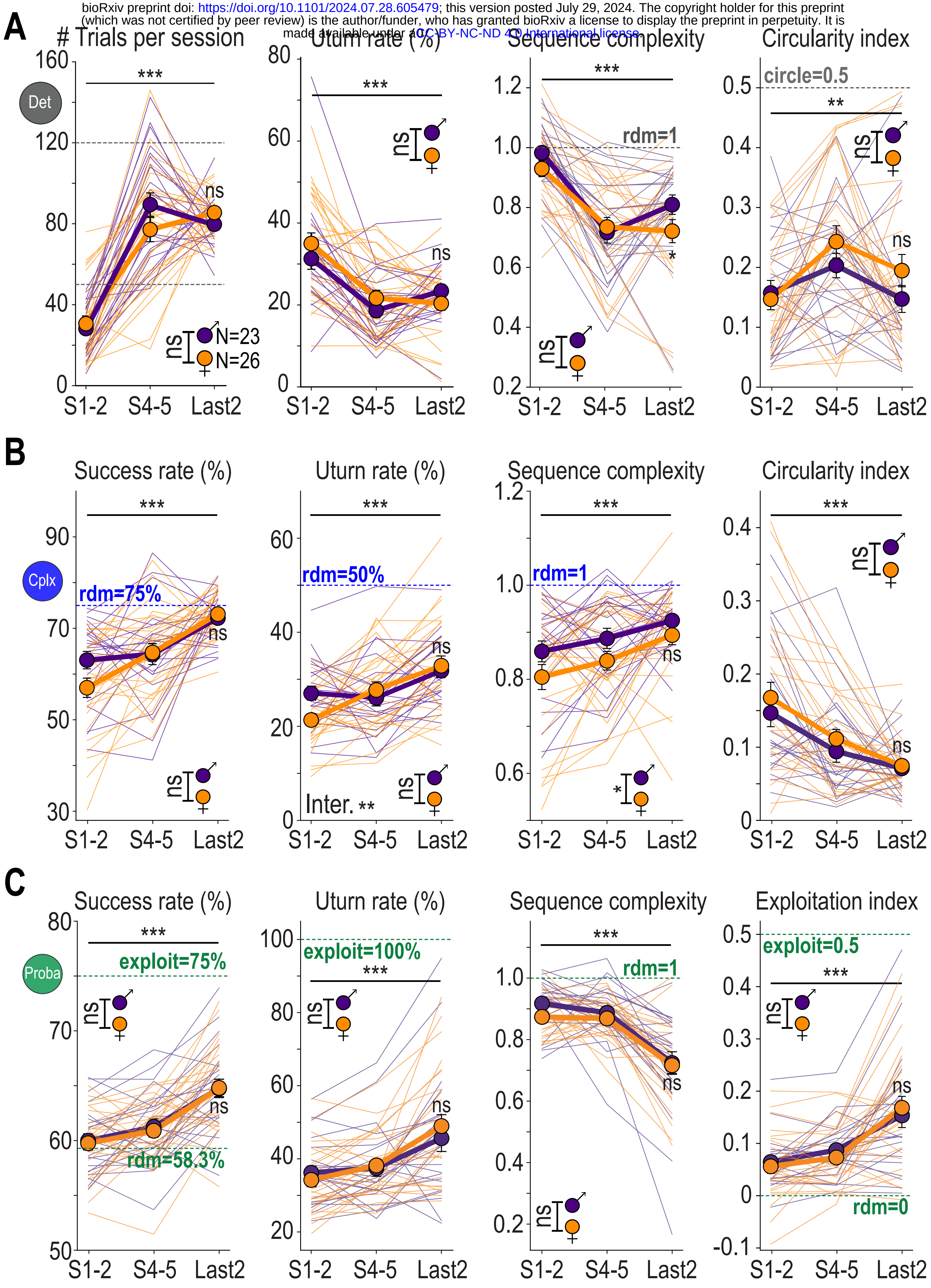
Reinforcement Learning (RL) models: We used Reinforcement Learning (RL) to compute Reward Prediction Errors (RPEs) from actual mice choice sequences and see how they match DA data. Before each trial, the agent contains a set of expected values for each possible action. As one of these actions is selected, it leads to either a reward ($V_{obtained}$ = 1) or an omission ($V_{obtained}$ = 0), then RPE is calculated as $V_{obtained}$ - $V_{expected}$, and a new expected value of this action is fed back into the agent's set for next trials. From both behavioural and photometry results, we hypothesised and tested three possible value representations in the bandit task. First, we proposed a simple, one-order representation "going to any target" or "performing any trial" to get a reward. In this case, all trials are similar, regardless of target or trajectory choices, and we simply compute and update $V_{expected}$ = { V_{Any} } at each trial. Second, a representation of internal directionality with a set of two actions and $V_{expected}$ = { V_{Fwd} ; V_{Uturn} }. In this case, RPEs are specific and computed separately for each of the two actions. Third, a spatial representation "going to target X" with a set of three actions and $V_{expected}$ = { V_{pA} ; V_{pB} ; V_{pC} }. Again, RPEs are computed for each target independently. Modelling the RPE values resulting from each of those three representations allowed us to compare them and determine which simulation better replicates DA data in each context. Initial $V_{expected}$ were set consistently with behavior in the task. For Det End, they were all set to 0.99. For both Cplx End and Proba End, they were set as mean success rate computed from the two previous sessions. For example, for a given mouse, initial V_{Uturn} to initiate the RL model with choice sequence from sessions 9-10 is the proportion of rewarded Uturn trials from sessions 7-8. Exception is for V_{p100} in Proba End where it was also set to 0.99. We arbitrarily tested several learning rates α = {0.001; 0.01; 0.05; 0.2; 0.4}. Results were consistent with experimental data for α = {0.01; 0.05; 0.2; 0.4}. Smaller α (0.001) led to convergence that was too slow considering mice number of trials provided to models, while larger α made convergence in Det too quick. In Fig 3 and Fig S6, α is set to 0.05. We next assumed that in our recordings, DA = $V_{obtained}$ + RPE, and tested which representation accounted most in the error component using GLMs on top of our RL-computed RPEs (taking as input variables $V_{obtained}$ = {1; 0} for rewards or omissions, and theoretical RPEs computed from Model 1, 2 and 3). Similarly, models were applied for each mouse in a given context, then w_i parameters were averaged among mice for each context, and the average was statistically compared to 0. Significant weight indicates that this variable explains part of DA variations. Finally, we extended this compilation of RL-computed RPE values and GLM to fit RPE weights to DA data across sessions and contexts (Fig 4 and Fig S7). In this case, we started RL models with mice choice sequences in Det Start with all $V_{expected}$ equal to zero (naive agents), computed corresponding RPEs and updated corresponding $V_{expected}$. Consistent with mice progressively learning and updating values across sessions and contexts, the final $V_{expected}$ of a given time-point became the initial $V_{expected}$ of the next time point. For instance, from Det Start to Det Mid (all $V_{expected}$ becoming closer to 1, but not at the same speed). Or from Cplx End to Proba Start ($V_{expected}$ of each target therefore starting to diverge). To allow for longitudinal comparisons, we next scaled (z-score) our data (both experimental DA and RL models-computed RPEs) on each time point, applied GLMs on each time point, and then compared the weights *i)* across sessions in a given context and at each transition between contexts, and *ii)* each of them regarding its difference with 0.

Figures and Statistics: Raw figures were plotted using Python custom codes (mostly MatPlotLib library). Graphics, typography and layout were formatted with Adobe Illustrator. All statistical analyses were computed using Python with Scipy library and custom programs. Results were most frequently plotted as individual data points and mean ± sem. The total number of observations in each group and the statistics used are indicated in figure legends and detailed statistics tables: unless specified, data points indicate the number of mice (N) on which the statistics were performed, and in some cases, they represent number of trials (n) either for one example session from one animal, or from all sessions of all animals in a given condition. Classical comparisons between means were performed using parametric tests (Student's t-test, or ANOVA for comparing more than two groups, when parameters followed a normal distribution (Shapiro test $P > 0.05$)), and non-parametric tests when the distribution was skewed (here, Wilcoxon or Mann-Whitney U for one/two samples and whether comparison is paired or not, or Kruskall-Wallis for more than two groups). More complex comparisons with several factors were performed using two-way or mixed ANOVA regardless of normal distribution for simplicity, with no major impact on results interpretation (see Fig S1, sex X session effects). Multiple comparisons were corrected using a sequentially rejective multiple test procedure (Holm). Linear regressions were assessed either with Pearson (parametric) or Spearman (non-parametric) tests. Probability distributions were compared using the Kolmogorov–Smirnov (KS) test. All statistical tests were two-sided. $p > 0.05$ was considered not to be statistically significant. In some cases, p > but close to 0.05 were indicated in the figure (see Tables of detailed statistics for more information).

Fluorescence immunohistochemistry: After completing the successive rules of the task, mice from the GRAB$_{DA}$ cohorts were euthanatized by IP injection of euthasol (0.1mL per 30g at 150mg/kg), immediately followed by paraformaldehyde (PFA) intra-cardiac perfusion, and brains were rapidly removed and post-fixed in 4% PFA for 2 to 4 days. Serial 60µm sections were cut with a vibratome (Leica). Immunohistochemistry was performed as follows: free-floating VTA and NAc brain sections were incubated for 1h at 4°C in a blocking solution of phosphate-buffered saline (PBS) containing 3% bovine serum albumin (BSA, Sigma A4503) and 0.2% Triton X-100, and then incubated overnight at 4 °C with *i)* a mouse anti-tyrosine hydroxylase primary antibody (TH, Sigma, T1299) at 1:500 dilution and *ii)* a chicken anti-eYFP primary antibody (Life technologies Molecular Probes, A- 6455) at 1:1000 dilution, both in PBS containing 1.5% BSA and 0.2% Triton X-100. The following day, sections were rinsed with PBS and then incubated for 3 h at 22–25 °C with *i)* Cy3-conjugated anti-mouse secondary antibody (Jackson ImmunoResearch, 715-165-150) at 1:500 dilution and *ii)* a goat anti-chicken AlexaFluor 488 secondary antibody (711-225-152, Jackson ImmunoResearch) at 1:1000 dilution, both in a solution of 1.5% BSA and 0.2% Triton X-100 in PBS. After three rinses in PBS, slices were wet-mounted using Prolong Gold Antifade Reagent with DAPI (Invitrogen, P36930). Microscopy was carried out with a fluorescent microscope Leica DMR, and images captured in gray level using MetaView software (Universal Imaging Corporation) and colored post-acquisition with ImageJ. Labeling for YFP in the NAc (along with satisfying signal during the task) allowed to confirm GRAB$_{DA}$ expression, and fiber implantation in the NAc lateral shell was also visually checked. Similar procedures were used to check for GCaMP7c and GCaMP6f expression in VTA

DA neurons. For GCaMP7c we used the same anti-TH and anti-eYFP antibodies as previously described. For GCaMP6f we used a sheep anti-TH primary antibody (AB-1542, Milipore) at 1:500 dilution coupled with a donkey anti-sheep secondary antibody (713-165-147, Jackson ImmunoResearch) at 1:500 dilution to highlight DA neurons, and simply used the virus-associated tdTomato to validate expression in the VTA and optic fiber implantation site. For MFB slices, 100 µm sections were performed and slices were directly visualized with visible light to check for ICSS electrode implantations.

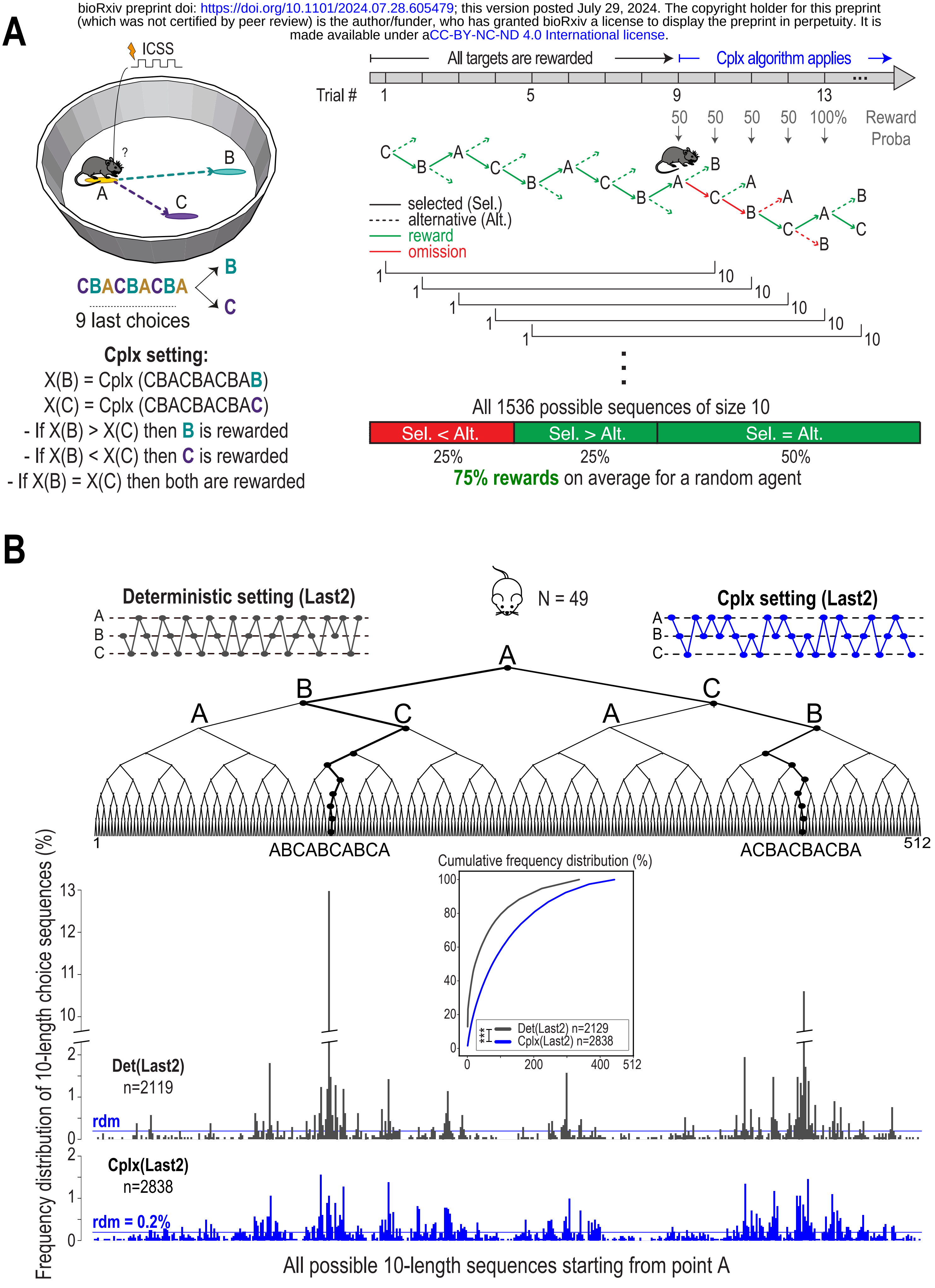
Statistics and Reproducibility: All experiments were replicated with success (several successive cohorts of mice)..



Supplementary figure 1

Fig. S1: Evolution of decision behaviour across sessions, with no major sex effects.

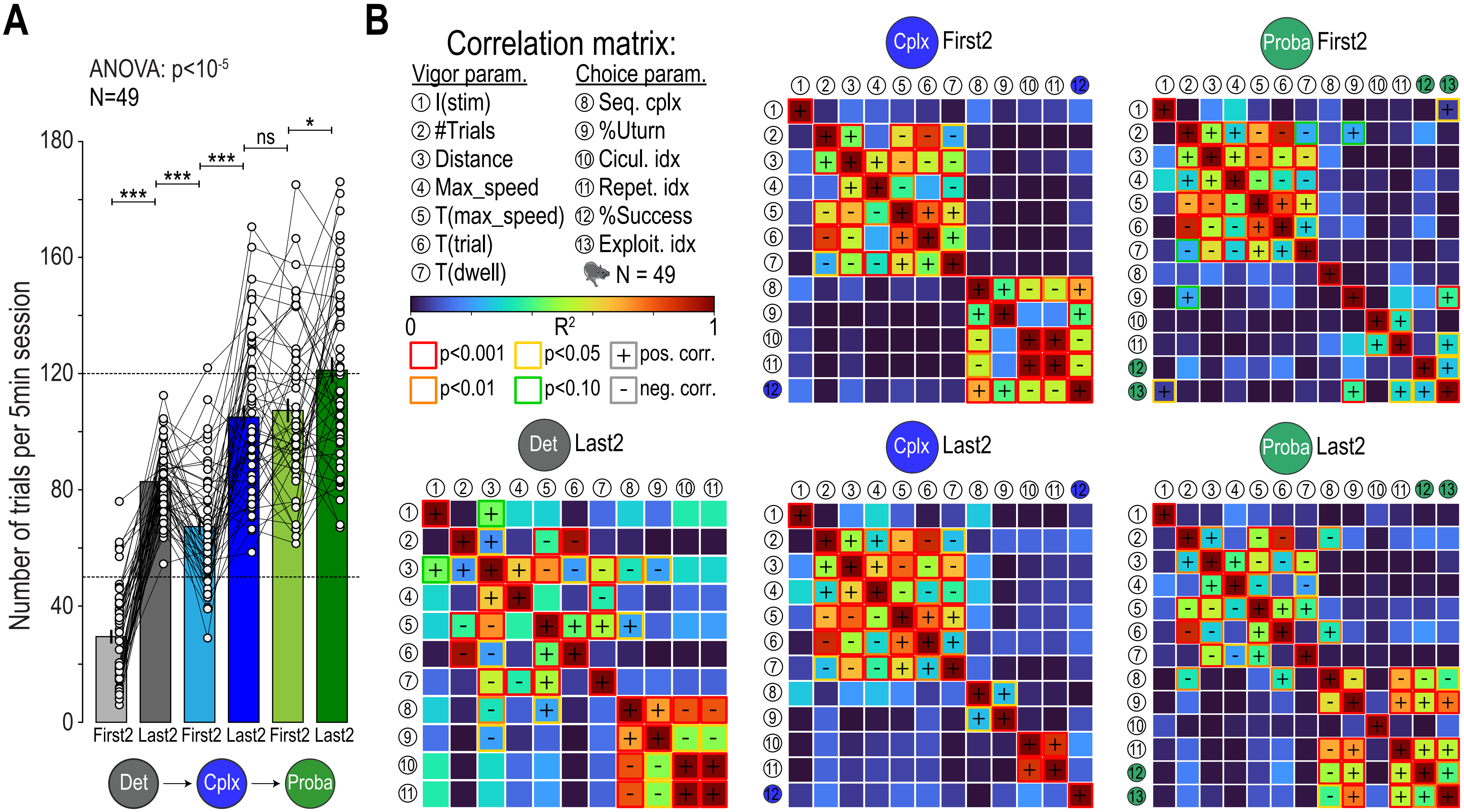
A. Decision parameters throughout Det sessions for males and females. Comparison of **(left)** the number of trials per session, **(middle-left)** the Uturn rate, **(middle-right)** the sequence complexity, and **(right)** the circularity index between sessions 1&2, sessions 4&5 and the last 2 sessions in male and female mice. In addition, we also compared the final states (Last2) between males and females. A fully circular mouse would have 0% Uturn, low seq. cplx and 0.5 circul. idx. **B. Same as in A) for Cplx sessions.** A mouse keeping circular strategy would have low success, 0% Uturn, low seq. cplx and 0.5 circul. idx. A random mouse would have 75% success, 50% Uturn, seq cplx = 1 and circul. idx = 0. **C. Same as in A) for Proba sessions.** An exploitative mouse would have 75% success, 100% Uturn, low seq. cplx and 0.5 exploit. idx. A random mouse would have 58.3% success, 50% Uturn, seq cplx = 1 and exploit. idx = 0. (Data are shown as individual points, and mean ±sem. N = 23 male and 26 female mice.)



Supplementary figure 2

Fig. S2: Additional information on the Cplx rule and mice sequence patterns.

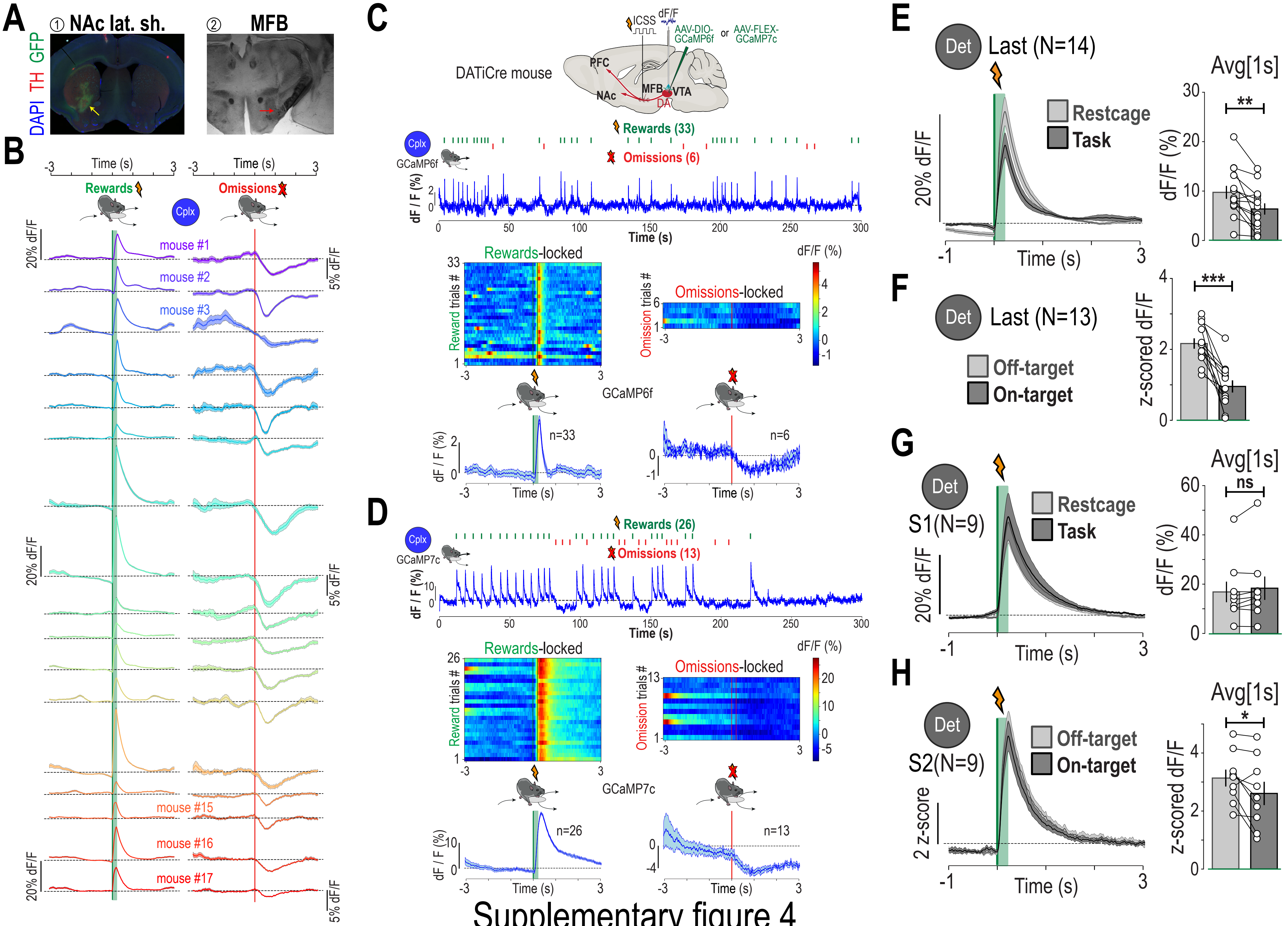
A. Detailed schematic representation of the Cplx rule. The first 9 trials of each session provide deterministic rewards (P=100%) to launch the Cplx algorithm, which then determines at each trial, in a sliding window, which target will lead to a reward by comparing the Lempel-Ziv grammatical complexity of the two potential sequences: 9 past choices + first remaining target VS. 9 past choices + second remaining target. The mouse will be rewarded only if it chooses the target that increases complexity. If both sequences have the same complexity, both targets will be rewarded **(see Methods)**. Taking all possible sequences of size 10 starting from one location, 75% of them are rewarded on the 10th trial. Therefore, a random agent exploring homogeneously this sequences tree will converge to 75% success rate. **B. Distribution of mice choice sequences of length 10 at the end of Det and Cplx.** Two distribution peaks (paths in the decision tree) appear in Det, corresponding to circling behavior (clockwise and counterclockwise), representing together roughly 25% of all produced sequences (among 512 possibilities). In Cplx, these peaks strongly reduce in size, in favor of more distributed visits of all possible sequences. **(Insert)** Cumulative distribution comparison between Det and Cplx (Last2 sessions for each rule). (In B, n is the total number of sequences of length 10, computed from sessions-wise mice successive choices, from N=49 mice both males and females).



Supplementary figure 3

Fig. S3: Motivation throughout the task and decoupling between vigor and choice parameters.

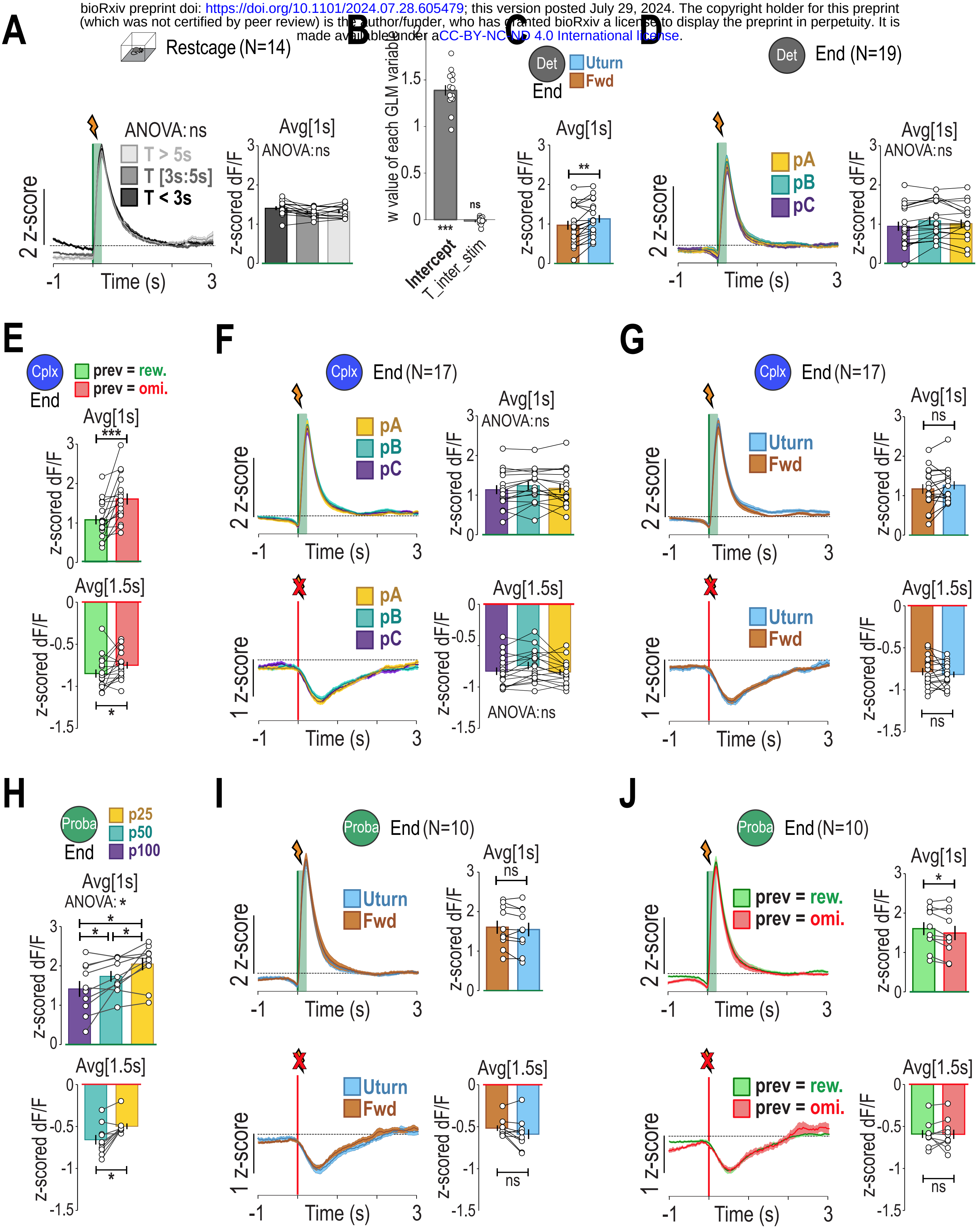
A. Evolution of motivation across contexts and sessions. Comparison of number of trials across contexts and sessions (mean of 2 sessions each time). **B. Correlation matrices between various vigor and choice parameters across mice in different contexts.** Parameters are computed for each mouse as the mean of 2 sessions (either First2 or Last2, for a given context). Each box represents the linear correlation between two parameters (Pearson for parametric, Spearman for non-parametric, each dot being a mouse). The filling color of each box represents the R^2 value. The frame color of each box represents the p-value (after Bonferroni correction). The warmer the color, the more those two parameters are significantly correlated. **(Left)** Last2 sessions of Det (11 parameters, x66 Bonferonni correction). **(Middle)** First2 and Last2 sessions of Cplx (12 parameters, x78 Bonferonni correction). **(Right)** First2 and Last2 sessions of Proba (13 parameters, x91 Bonferonni correction). (In A, data are shown as individual points, and mean ±sem. In B, only R^2 and corrected p-values are shown with color code. Individual data are available upon request. N is always the number of mice.)



Supplementary figure 4

Fig. S4: DA fiber photometry signals in various configurations.

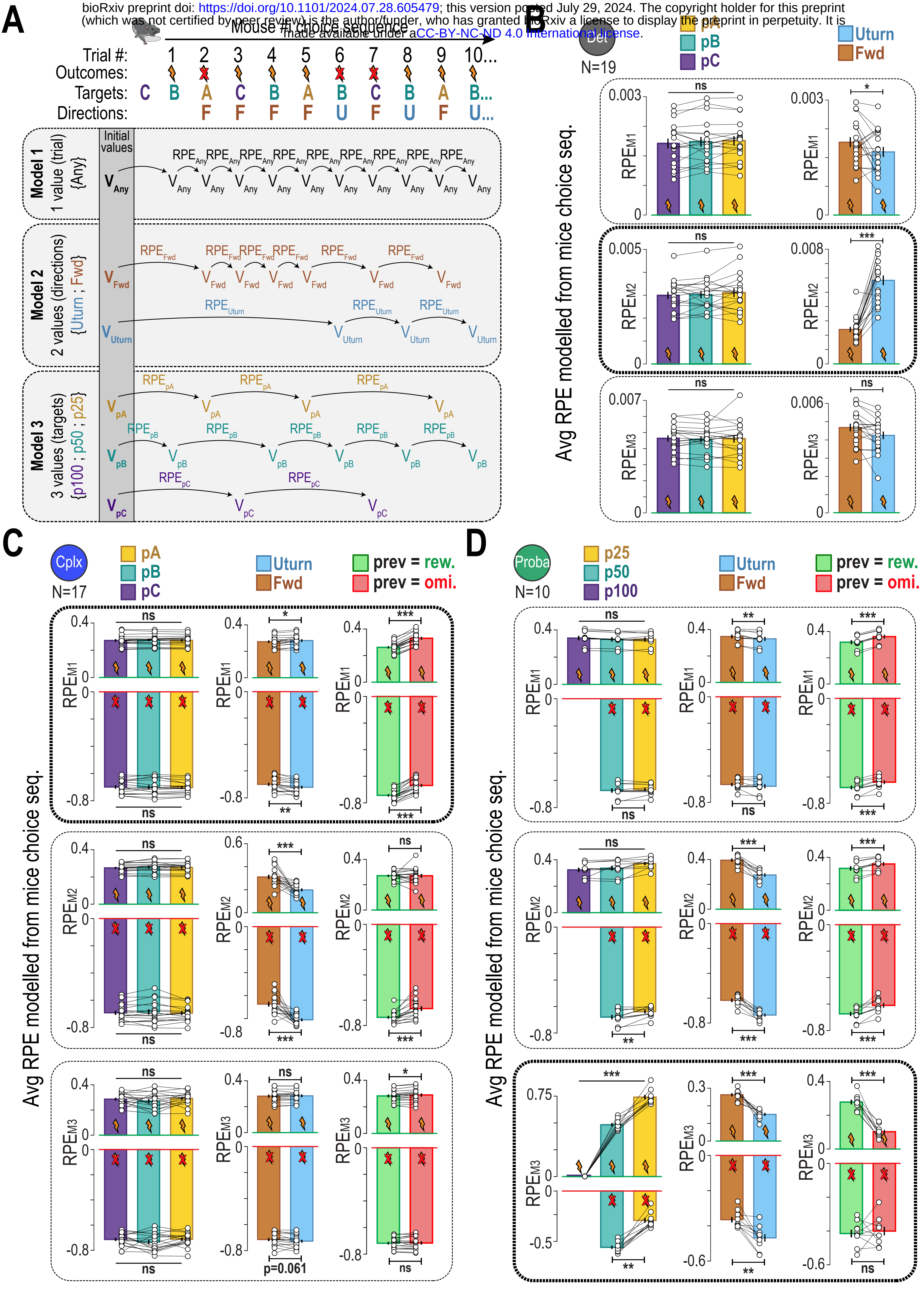
A. NAc and MFB slices immunohistochemistry. Post-hoc verification of optic fiber implant and Grab$_{DA}$ virus expression in the NAc lateral shell (left), and stimulation electrode implant in the MFB (right). **B. Individual mice NAc DA release for rewards and omissions in Cplx.** Each line and colour is an individual mouse, averaged for all trials during last Cplx session, in [-3s:3s] time window locked on location entry. Every single mouse included in the results displayed reward-induced peaks and omission-induced dips of DA release significantly different from zero (dashed black lines). **C-D. DA cell activity using GCaMP fiber photometry.** DATiCre mice were injected with an AAV to express either GCaMP6f or GCaMP7c in VTA DA neurons, implanted with an optic fiber in the VTA, and stimulation electrode in the MFB, to assess DA neuron activity in the task. **C. GCaMP6f.** Using similar experimental procedures and signal analyses in the Cplx context, calcium dynamics of VTA DA neurons show similar reward-induced peaks and omission-induced dips than NAc lateral shell DA release, in this case with faster kinetics for peaks, and smaller signal amplitudes (worse signal-to-noise ratio) for both peaks and dips. **D. GCaMP7c.** Same as B for GCaMP7c, with slower kinetics for peaks, and greater signal amplitudes (better signal-to-noise ratio) for both peaks and dips. **E. DA response to expected (Task) vs unexpected (Restcage) rewards in Det Last session.** Comparison between Task and Restcage ICSS (same session, same current intensity). **F. DA response to expected (On-target) vs unexpected (Off-target) rewards in Det Last session.** Individual data corresponding to Fig2.E. Comparison between On-target and Off-target ICSS (same session, same current intensity). **G. DA response to Task vs Restcage rewards in Det first (S1) session.** Same as D but during first session (S1) of conditioning in Det. **H. DA response to On-target vs Off-target rewards in Det second (S2) session.** Same as E but during second session (S2) of conditioning in Det. (In B, C, D, curves are shown as mean ±sem for a single session, n is the number of reward or omission trials in this session. In E, G, H, curves are shown as mean ±sem for session-wise average of several mice, N is the number of mice in each condition. In E, F, G, H, Bar plots are shown as mean ±sem, in addition to individual data points.)



Supplementary figure 5

Fig. S5: Additional analyses of NAc DA release regarding task behavioural features.

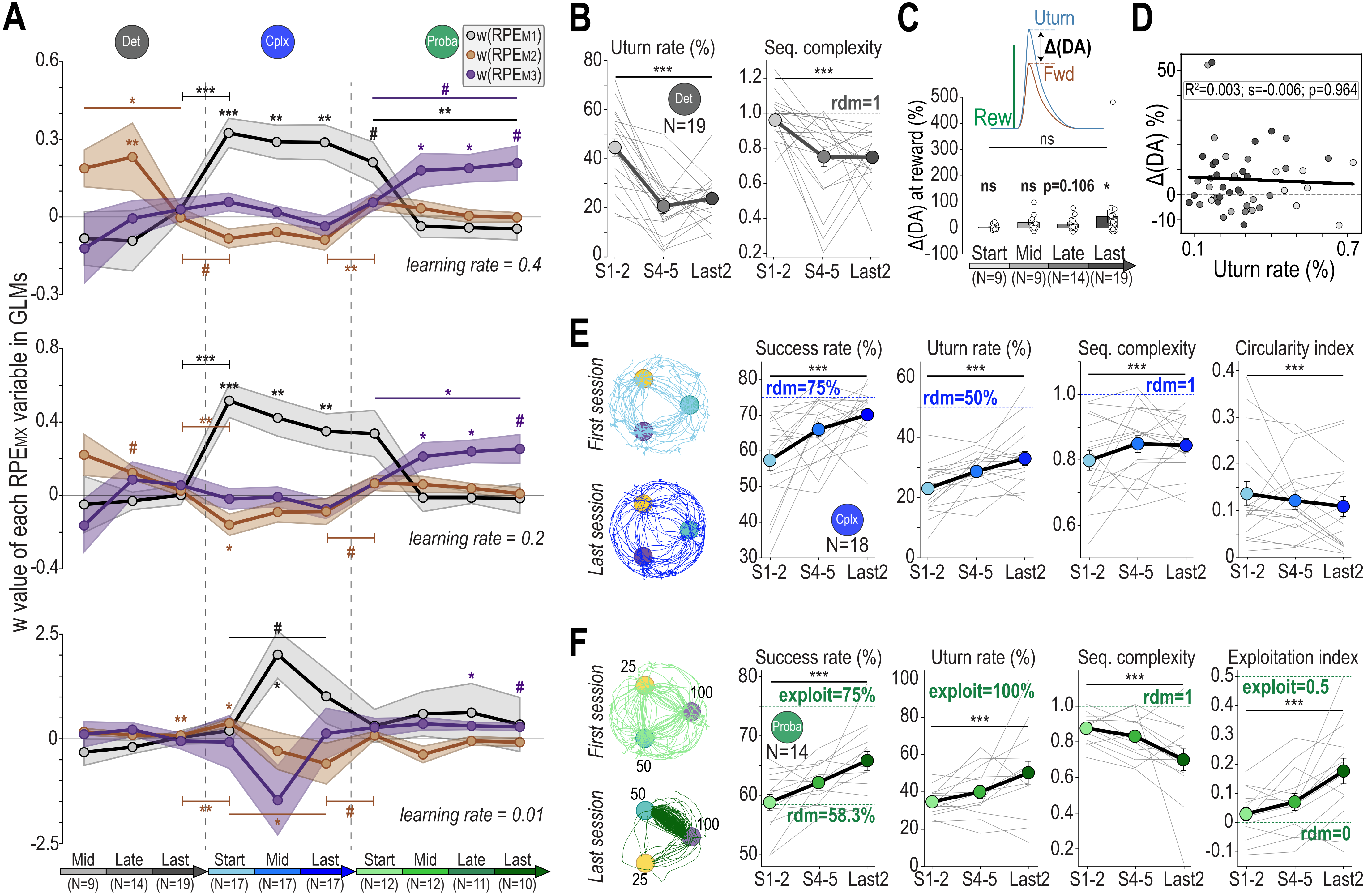
A. GLM in Restcage the same day as Last Det. Weight value of each variable compared to zero. **B. Inter-stimulation interval effect on DA transients in Restcage the same day as Last Det.** Comparison between short (<3s), mid ([3s:5s]) and long (>5s) intervals. **C. Trajectory effect on DA transients in Det End.** Individual data corresponding to Fig2.J. Comparison between Fwd and Uturn. **D. Target effect on DA transients in Det End.** Comparison between pA, pB and pC. **E. Previous outcome effect on DA transients in Cplx End.** Individual data corresponding to Fig2.K. **(Top)** Reward peak comparison between previous reward and previous omission. **(Bottom)** Omission dip comparison between previous reward and previous omission. **F. Target effect on DA transients in Cplx End. (Top)** Reward peak comparison between pA, pB and pC. **(Bottom)** Omission dip comparison between pA, pB and pC. **G. Trajectory effect on DA transients in Cplx End. (Top)** Reward peak comparison between Uturn and Fwd. **(Bottom)** Omission dip comparison between Uturn and Fwd. **H. Target effect on DA transients in Proba End.** Individual data corresponding to Fig2.L. **(Top)** Reward peak comparison between p100, p50 and p25. **(Bottom)** Omission dip comparison between p50 and p25. **I. Trajectory effect on DA transients in Proba End. (Top)** Reward peak comparison between Uturn and Fwd. **(Bottom)** Omission dip comparison between Uturn and Fwd. **I. Previous outcome effect on DA transients in Proba End. (Top)** Reward peak comparison between previous reward and previous omission. **(Bottom)** Omission dip comparison between previous reward and previous omission. (In A, B, C, D, E, F, G, H, I, J Bar plots are shown as mean ±sem, in addition to individual data points. In A, D, F, G, I, J, signal curves are shown as mean ±sem. N is always the number of mice in each context.)



Supplementary figure 6

Fig. S6: Additional information on the three RL models, comparison of computed RPEs in various behavioural scenarios, and results in Proba Change context.

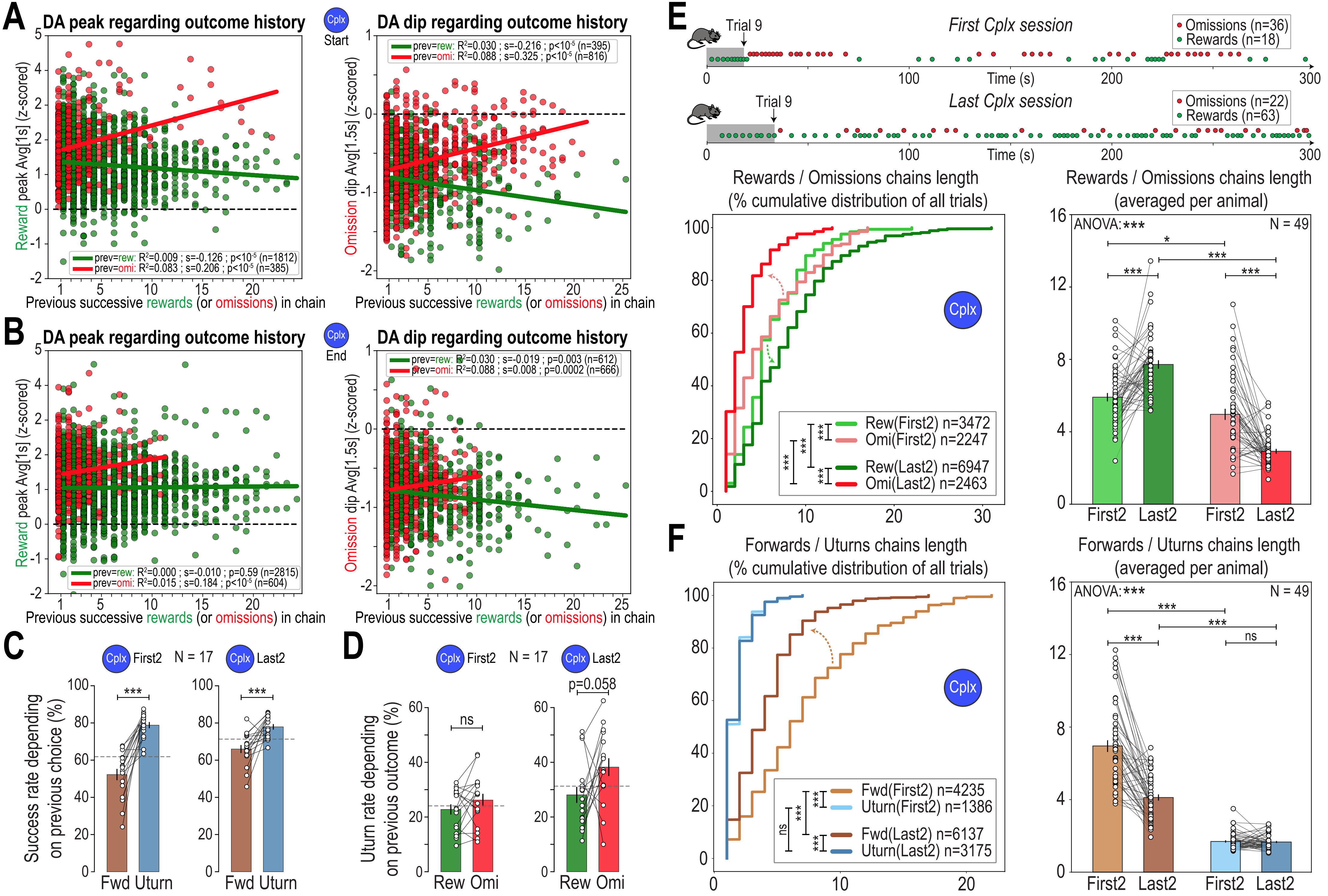
A. Detailed schematic of RL modelling for each of the three models. From actual mice choice sequences we applied RL models and computed corresponding RPEs. The first model consists in single value representation "going to any target" or "performing any trial" to get a reward, where we simply compute $V_{expected}$ = { V_{Any} } and RPE_{Any} at each trial. The second model consists in two value representations depending on chosen trajectory $V_{expected}$ = { V_{Fwd} ; V_{Uturn} }. In this case, RPE_{Uturn} and RPE_{Fwd} are specific and computed separately for each of those two actions. The third model consists in three value representations depending on chosen target $V_{expected}$ = { V_{pA} ; V_{pB} ; V_{pC} }. Again, RPE_{pA}, RPE_{pB} and RPE_{pC} are computed for each target independently. **B-C-D. For each model, computed RPEs were averaged over mice sessions in the same scenarios used to characterise DA responses (regarding target, trajectory, and previous outcome).** The model that qualitatively reproduces best DA responses in all scenarios in given context is supposed to be the best value representation that mice are using in this context. **B. End Det context. Top:** Average M1-computed RPE comparison between targets, and trajectories. **Center:** Same for M2 (same as Fig3.E). **Bottom:** Same for M3. **C. End Cplx context. Top:** Average M1-computed RPE comparison between targets, trajectories and previous outcome (same as Fig3.F). **Center:** Same for M2. **Bottom:** Same for M3. **D. End Proba context. Top:** Average M1-computed RPE comparison between targets, trajectories and previous outcome. **Center:** Same for M2. **Bottom:** Same for M3 (same as Fig3.G). (In B, C, D Bar plots are shown as mean ±sem, in addition to individual data points. In G, signal curves are shown as mean ±sem. N is always the number of mice in each context.)



Supplementary figure 7

Fig. S7: Evolution of Model weights, DA transients and strategy parameters across rules and sessions.

A. RL modelling and GLM fitting DA data with computed RPEs across sessions and contexts. Same as Fig4A with varying learning rates. **Top:** For learning rate α=0.4, evolution of RPE_{M1} (black), RPE_{M2} (brown) and RPE_{M3} (purple) weights over time and multiple comparisons of each time point with zero. **Middle:** Same for learning rate α=0.2. **Bottom:** Same for learning rate α=0.01. **B. Evolution of choice parameters across Det sessions.** Comparison of **(left)** Uturn rate and **(right)** sequence complexity between sessions 1&2, sessions 4&5 and last 2 sessions in Grab-DA mice. **C. Comparison of ΔDA(directions) across Det sessions. Top:** ΔDA is computed for each mouse as the relative difference Δ = (Uturn—Fwd) / Fwd. **Bottom:** Comparison of ΔDA between Start, Mid, Late and Last sessions, and multiple comparisons of each time-point with zero. **D. Linear regressions between ΔDA(directions) and Uturn in Det. Top:** Reward ΔDA regarding Uturn rate of each mouse at each time point (light grey Start => dark grey Last). **E. Evolution of choice parameters across Cplx sessions. Left:** Example trajectories of first (cyan) and last (blue) Cplx sessions. Comparison of **(middle-left)** Success rate, **(middle)** Uturn rate, **(middle-right)** sequence complexity and **(right)** circularity index between sessions 1&2, sessions 4&5 and last 2 sessions in Grab-DA mice. **F. Evolution of choice parameters across Proba sessions. Left:** Example trajectories of first (light green) and last (dark green) Proba sessions. Comparison of **(middle-left)** Success rate, **(middle)** Uturn rate, **(middle-right)** sequence complexity and **(right)** exploitation index between sessions 1&2, sessions 4&5 and last 2 sessions in Grab-DA mice. (In B, C, E, F, data are shown as mean ±sem, in addition to individual data points. In D, each data point is one animal at one time point. signal curves are shown as mean ±sem. In A, data are shown as mean ±sem for clarity. Individual data points are available upon requests. Due to multiple corrections generating dilutions in p-values, **#** symbol has been used in the figure to highlight p<0.12 after correction. N is always the number of mice in each context.)



Supplementary figure 8

Fig. S8: Additional analyses of DA transients and choice behavior in Cplx.

A. Linear regressions of DA transients depending on the number of successive previous rewards or omissions in chains in Cplx Start. Left: Reward-induced DA peak amplitudes regarding length of successive previous rewards chains (green) or omissions chains (red). **Right:** Same for omission-induced DA dip amplitudes regarding length of successive previous rewards chains (green) or omissions chains (red). **B. Same for Cplx End. Left:** Reward-induced DA peak amplitudes regarding length of successive previous rewards chains (green) or omissions chains (red). **Right:** Same for omission-induced DA dips amplitude regarding length of successive previous rewards chains (green) or omissions chains (red). **C. Success rate depending on previous Uturn/Fwd choice in Cplx. Left:** First 2 sessions. **Right:** Last 2 sessions. **D. Uturn rate depending on previous outcome in Cplx. Left:** First 2 sessions. **Right:** Last 2 sessions. **E. Analysis of chains of successive rewards and omissions in Cplx. Top:** In early Cplx, mice tend to keep repeating circular patterns and therefore get long series of omissions. In late Cplx, omissions are regularly distributed, generating smaller chains, as expected from a random agent. **Bottom-left:** Cumulative distribution of reward and omission chain lengths during first 2 and last 2 Cplx sessions. **Bottom-right:** Average chain lengths per mouse. **E. Same for chains of successive forwards and Uturns. Left:** Cumulative distribution of forward and Uturn chains length during first 2 and last 2 Cplx sessions. **Right:** Average chains length per mouse. For regressions in A, B, each dot is a trial of one mouse. In C, D, E, F, Bar plots are shown as mean ±sem, in addition to individual data points. In E, F, cumulative distribution are computed for all trials of all mice together. n is always the number of trials, and N the number of mice, in each context.)

Tables of detailed statistics for figures 1-4 and supp 1-8:

Figure 1

Panel	Comparison	Test type	p-values	Corrections
E (top, right)	%Success across tasks, all mice (N=49), Cplx vs Proba	Student paired t-test	**$p<10^{-5}$**	
E (bottom, left)	Seq cplx across tasks, all mice (N=49), Det vs Cplx vs Proba	one-way ANOVA	**$p<10^{-5}$**	
	Post-hoc, Det vs Cplx	Wilcoxon (paired)	**$p<10^{-5}$**	**Holm (x3) : $p<10^{-5}$**
	Post-hoc, Cplx vs Proba	Wilcoxon (paired)	**$p<10^{-5}$**	**Holm (x3) : $p<10^{-5}$**
	Post-hoc, Det vs Proba	Wilcoxon (paired)	p=0.2121	Holm (x3) : p=0.2121
E (bottom, right)	%Uturns across tasks, all mice (N=49), Det vs Cplx vs Proba	one-way ANOVA	**$p<10^{-5}$**	
	Post-hoc, Det vs Cplx	Wilcoxon (paired)	**$p<10^{-5}$**	**Holm (x3) : $p<10^{-5}$**
	Post-hoc, Cplx vs Proba	Wilcoxon (paired)	**$p<10^{-5}$**	**Holm (x3) : $p<10^{-5}$**
	Post-hoc, Det vs Proba	Wilcoxon (paired)	**$p<10^{-5}$**	**Holm (x3) : $p<10^{-5}$**
F (top, left)	%Visits in Det, all mice (N=49), pA vs pB vs pC (N=3)	one-way ANOVA (target effect)	Target effect: p=0.1796	
F (top, right)	Gamble %Pref in Det, all mice (N=49), gA vs gB vs gC (N=3)	one-way ANOVA (gamble effect)	Gamble effect: p=0.9029	
F (middle, left)	%Visits in Cplx, all mice (N=49), pA vs pB vs pC (N=3)	one-way ANOVA (target effect)	Target effect: p=0.9786	
F (middle, right)	Gamble %Pref in Cplx, all mice (N=49), gA vs gB vs gC (N=3)	one-way ANOVA (gamble effect)	Gamble effect: p=0.9516	
F (bottom, left)	%Visits in Proba, all mice (N=49), p100 vs p50 vs p25 (N=3)	one-way ANOVA (target effect)	**Target effect: $p<10^{-5}$**	
F (bottom, right)	Gamble %Pref in Proba, all mice (N=49), g100 vs g50 vs g25 (N=3)	one-way ANOVA (gamble effect)	**Gamble effect: $p<10^{-5}$**	

Figure 2

Panel	Comparison	Test type	p-values	Corrections
C (bottom, left)	Post-reward 1s-avg dF/F (n=47 trials) for one single session, vs 0	one sample Student t-test	**$p<10^{-5}$**	
C (bottom, right)	Post-omission 1.5s-avg dF/F (n=32 trials) for one single session, vs 0	one sample Student t-test	**$p<10^{-5}$**	
D	Post-reward 1s-avg dF/F for all mice End sessions : Restcage (n=988) vs Det (n=2288) vs Cplx (n=3150 trials) vs Proba (n=1704)	Kolmogorov-Smirnov (distribution)	**Restcage vs Det : $p<10^{-5}$** **Restcage vs Cplx : $p<10^{-5}$** **Restcage vs Proba : $p<10^{-5}$** **Det vs Cplx : $p<10^{-5}$** **Det vs Proba : $p<10^{-5}$** **Cplx vs Proba : $p<10^{-5}$**	**Holm (x6) : all $p<10^{-5}$**
	Post-omission 1.5s-avg dF/F for all mice End sessions, Cplx (n=1107 trials) vs Proba (n=845)	Kolmogorov-Smirnov (distribution)	**$p<10^{-5}$**	
E (right)	Post-ICSS avg per mouse (N=13), Expected (on-target) vs Unexpected (off-target)	Student paired t-test	**$p<10^{-5}$**	
G	Det End GLM : Intercept weight vs 0 (N=19)	one sample Student t-test	**$p<10^{-5}$**	**Holm (x3) : $p<10^{-5}$**
	Det End GLM : Uturn weight vs 0 (N=19)	one sample Student t-test	**p=0.0007**	**Holm (x3) : p=0.0013**
	Det End GLM : Target weight vs 0 (N=19)	one sample Student t-test	p=0.1171	Holm (x3) : p=0.1171
H	Cplx End GLM : Intercept weight vs 0 (N=17)	one sample Student t-test	p=0.0672	Holm (x6) : p=0.2016
	Cplx End GLM : Reward weight vs 0 (N=17)	one sample Student t-test	**$p<10^{-5}$**	**Holm (x6) : $p<10^{-5}$**
	Cplx End GLM : Omission weight vs 0 (N=17)	one sample Student t-test	**$p<10^{-5}$**	**Holm (x6) : $p<10^{-5}$**
	Cplx End GLM : Uturn weight vs 0 (N=17)	one sample Student t-test	p=0.3264	Holm (x6) : p=0.3264
	Cplx End GLM : Target weight vs 0 (N=17)	one sample Student t-test	p=0.0875	Holm (x6) : p=0.2016
	Cplx End GLM : Previous omission weight vs 0 (N=17)	one sample Wilcoxon	**p=0.00002**	**Holm (x6) : p=0.0002**
I	Proba End GLM : Intercept weight vs 0 (N=10)	one sample Student t-test	**p=0.00001**	**Holm (x6) : p=0.00005**
	Proba End GLM : Reward weight vs 0 (N=10)	one sample Wilcoxon	**p=0.0020**	**Holm (x6) : p=0.0078**
	Proba End GLM : Omission weight vs 0 (N=10)	one sample Student t-test	**$p<10^{-5}$**	**Holm (x6) : $p<10^{-5}$**
	Proba End GLM : Uturn weight vs 0 (N=10)	one sample Student t-test	p=0.1615	Holm (x6) : p=0.3229
	Proba End GLM : Target_proba weight vs 0 (N=10)	one sample Student t-test	**p=0.0090**	**Holm (x6) : p=0.0269**
	Proba End GLM : Previous omission weight vs 0 (N=10)	one sample Student t-test	p=0.4280	Holm (x6) : p=0.4280
J (right)	Det End post-reward dF/F avg per mouse (N=19) : Uturn vs Forward	Student paired t-test	**p=0.0012**	
K (left)	Cplx End post-reward dF/F avg per mouse (N=17) : previous=rew vs previous=omi	Student paired t-test	**p=0.0003**	
K (right)	Cplx End post-omission dF/F avg per mouse (N=17) : previous=rew vs previous=omi	Student paired t-test	**p=0.0357**	
L (left)	Proba End post-reward dF/F avg per mouse (N=10) : p100 vs p50 vs p25	one-way ANOVA (target effect)	**p=0.0364**	
	Post-hoc, p100 vs p50	Student paired t-test	**p=0.0171**	**Holm (x3) : p=0.0282**
	Post-hoc, p50 vs p25	Student paired t-test	**p=0.0141**	**Holm (x3) : p=0.0282**
	Post-hoc, p100 vs p25	Student paired t-test	**p=0.0079**	**Holm (x3) : p=0.0237**
L (right)	Proba End post-omission dF/F avg per mouse (N=10) : p50 vs p25	Student paired t-test	**p=0.0173**	

Figure 3

Panel	Comparison	Test type	p-values	Corrections
B	Det End GLM on models RPE avg per mouse (N=19) : Intercept weight vs 0	one sample Student t-test	**p=0.00002**	**Holm (x4) : p=0.0002**
	Det End GLM on models RPE avg per mouse (N=19) : RPE(M1) weight vs 0	one sample Student t-test	p=0.9842	Holm (x4) : p=1
	Det End GLM on models RPE avg per mouse (N=19) : RPE(M2) weight vs 0	one sample Wilcoxon	**p=0.0024**	**Holm (x4) : p=0.0072**
	Det End GLM on models RPE avg per mouse (N=19) : RPE(M3) weight vs 0	one sample Wilcoxon	p=0.5153	Holm (x4) : p=1
C	Cplx End GLM on models RPE avg per mouse (N=17) : Intercept weight vs 0	one sample Student t-test	p=0.2690	Holm (x5) : p=0.5380
	Cplx End GLM on models RPE avg per mouse (N=17) : V(obtained) weight vs 0	one sample Student t-test	**p=0.0053**	**Holm (x5) : p=0.0221**
	Cplx End GLM on models RPE avg per mouse (N=17) : RPE(M1) weight vs 0	one sample Student t-test	**p=0.0044**	**Holm (x5) : p=0.0221**
	Cplx End GLM on models RPE avg per mouse (N=17) : RPE(M2) weight vs 0	one sample Student t-test	p=0.1591	Holm (x5) : p=0.4773
	Cplx End GLM on models RPE avg per mouse (N=17) : RPE(M3) weight vs 0	one sample Student t-test	p=0.4564	Holm (x5) : p=0.5380
D	Proba End GLM on models RPE avg per mouse (N=10) : Intercept weight vs 0	one sample Student t-test	p=0.9262	Holm (x5) : p=1
	Proba End GLM on models RPE avg per mouse (N=10) : V(obtained) weight vs 0	one sample Student t-test	**p=0.0334**	**Holm (x5) : p=0.1337**
	Proba End GLM on models RPE avg per mouse (N=10) : RPE(M1) weight vs 0	one sample Student t-test	p=0.5484	Holm (x5) : p=1
	Proba End GLM on models RPE avg per mouse (N=10) : RPE(M2) weight vs 0	one sample Student t-test	p=0.9562	Holm (x5) : p=1
	Proba End GLM on models RPE avg per mouse (N=10) : RPE(M3) weight vs 0	one sample Student t-test	**p=0.0111**	**Holm (x5) : p=0.0556**
E	Det End RPE(M2) avg per mouse (N=19) : Uturn vs Forward	Wilcoxon (paired)	**p=0.00002**	
F (right, top)	Cplx End post-reward RPE(M1) avg per mouse (N=17) : previous=reward vs omission	Student paired t-test	**$p<10^{-5}$**	
F (right, bottom)	Cplx End post-omission RPE(M1) avg per mouse (N=17) : previous=reward vs omission	Student paired t-test	**$p<10^{-5}$**	
G (right, top)	Proba End post-reward RPE(M3) avg per mouse (N=10) : p100 vs p50 vs p25	Kruskall-Wallis	**$p<10^{-5}$**	
G (right, bottom)	Proba End post-omission RPE(M3) avg per mouse (N=10) : p50 vs p25	Wilcoxon (paired) t-test	**p=0.0019**	
H (left, bottom)	Proba Change post-omission dF/F avg per mouse (N=6) : p100=>50 vs p50 vs p25	Kruskall-Wallis	**p=0.0013**	
	Post-hoc, p100=>50 vs p50	Wilcoxon	**p=0.0313**	**Holm (x3) : p=0.0625**
	Post-hoc, p50 vs p25	Student paired t-test	**p=0.0060**	**Holm (x3) : p=0.0181**
	Post-hoc, p100=>50 vs p25	Wilcoxon	**p=0.0313**	**Holm (x3) : p=0.0625**
H (right)	Proba Change GLM : Intercept weight vs 0 (N=6)	one sample Student t-test	**p=0.0021**	**p=0.0082**
	Proba Change GLM : Reward weight vs 0 (N=6)	one sample Student t-test	**p=0.0008**	**p=0.0048**
	Proba Change GLM : Omission weight vs 0 (N=6)	one sample Student t-test	**p=0.0006**	**p=0.0043**
	Proba Change GLM : Uturn weight vs 0 (N=6)	Wilcoxon	p=0.5625	p=0.5625
	Proba Change GLM : Target_proba_old weight vs 0 (N=6)	one sample Student t-test	**p=0.0011**	**p=0.0056**
	Proba Change GLM : Target_proba_new weight vs 0 (N=6)	one sample Student t-test	p=0.1334	p=0.2668
	Proba End GLM : Previous omission weight vs 0 (N=6)	one sample Student t-test	p=0.0589	p=0.1768

Figure 4

Panel	Comparison	Test type	p-values	Corrections
A (top)	GLM RPE(M1) weight across Det	one-way ANOVA	p=0.7950	
	GLM RPE(M1) weight : End Det vs Start Cplx	Student unpaired t-test	**p=0.00006**	
	GLM RPE(M1) weight across Cplx	one-way ANOVA	p=0.2407	
	GLM RPE(M1) weight : End Cplx vs Start Proba	Student unpaired t-test	p=0.5875	
	GLM RPE(M1) weight across Proba	Kruskall-Wallis	p=0.8552	
	GLM RPE(M1) weight vs 0 : Det Mid	one sample Student t-test	p=0.6810	Holm (x10) : p=1
	GLM RPE(M1) weight vs 0 : Det Late	one sample Student t-test	p=0.8201	Holm (x10) : p=1
	GLM RPE(M1) weight vs 0 : Det Last	one sample Student t-test	p=0.4648	Holm (x10) : p=1
	GLM RPE(M1) weight vs 0 : Cplx Start	one sample Student t-test	**p=0.00002**	**Holm (x10) : p=0.0002**
	GLM RPE(M1) weight vs 0 : Cplx Mid	one sample Student t-test	**p=0.00006**	**Holm (x10) : p=0.0006**
	GLM RPE(M1) weight vs 0 : Cplx Last	one sample Student t-test	**p=0.0094**	**Holm (x10) : p=0.0754**
	GLM RPE(M1) weight vs 0 : Proba Start	one sample Student t-test	**p=0.0325**	Holm (x10) : p=0.2275
	GLM RPE(M1) weight vs 0 : Proba Mid	one sample Wilcoxon	p=0.1010	Holm (x10) : p=0.6592
	GLM RPE(M1) weight vs 0 : Proba Late	one sample Student t-test	p=0.3017	Holm (x10) : p=1
	GLM RPE(M1) weight vs 0 : Proba Last	one sample Student t-test	p=0.5886	Holm (x10) : p=1
	GLM RPE(M2) weight across Det	one-way ANOVA	p=0.5767	
	GLM RPE(M2) weight : End Det vs Start Cplx	Mann-Whitney U test (unpaired)	**p=0.00001**	
	GLM RPE(M2) weight across Cplx	Kruskall-Wallis	p=0.3804	
	GLM RPE(M2) weight : End Cplx vs Start Proba	Student unpaired t-test	p=0.3013	
	GLM RPE(M2) weight across Proba	one-way ANOVA	p=0.9336	
	GLM RPE(M2) weight vs 0 : Det Mid	one sample Student t-test	p=0.1612	Holm (x10) : p=1
	GLM RPE(M2) weight vs 0 : Det Late	one sample Student t-test	p=0.2922	Holm (x10) : p=1
	GLM RPE(M2) weight vs 0 : Det Last	one sample Student t-test	**p=0.0130**	**Holm (x10) : p=0.1168**
	GLM RPE(M2) weight vs 0 : Cplx Start	one sample Wilcoxon	**p=0.0002**	**Holm (x10) : p=0.0021**
	GLM RPE(M2) weight vs 0 : Cplx Mid	one sample Student t-test	**p=0.0398**	Holm (x10) : p=0.3187
	GLM RPE(M2) weight vs 0 : Cplx Last	one sample Student t-test	p=0.2023	Holm (x10) : p=1
	GLM RPE(M2) weight vs 0 : Proba Start	one sample Student t-test	p=0.9793	Holm (x10) : p=1
	GLM RPE(M2) weight vs 0 : Proba Mid	one sample Student t-test	p=0.8436	Holm (x10) : p=1
	GLM RPE(M2) weight vs 0 : Proba Late	one sample Student t-test	p=0.3895	Holm (x10) : p=1
	GLM RPE(M2) weight vs 0 : Proba Last	one sample Student t-test	p=0.8098	Holm (x10) : p=1
	GLM RPE(M3) weight across Det	one-way ANOVA	p=0.4364	
	GLM RPE(M3) weight : End Det vs Start Cplx	Student unpaired t-test	**p=0.0695**	
	GLM RPE(M3) weight across Cplx	Kruskall-Wallis	p=0.1509	
	GLM RPE(M3) weight : End Cplx vs Start Proba	Mann-Whitney U test (unpaired)	p=0.2406	
	GLM RPE(M3) weight across Proba	Kruskall-Wallis	**p=0.1157**	
	GLM RPE(M3) weight vs 0 : Det Mid	one sample Student t-test	p=0.3224	Holm (x10) : p=1
	GLM RPE(M3) weight vs 0 : Det Late	one sample Student t-test	p=0.6185	Holm (x10) : p=1
	GLM RPE(M3) weight vs 0 : Det Last	one sample Student t-test	p=0.6133	Holm (x10) : p=1
	GLM RPE(M3) weight vs 0 : Cplx Start	one sample Student t-test	**p=0.0068**	**Holm (x10) : p=0.0541**
	GLM RPE(M3) weight vs 0 : Cplx Mid	one sample Wilcoxon	**p=0.0202**	Holm (x10) : p=0.1210
	GLM RPE(M3) weight vs 0 : Cplx Last	one sample Wilcoxon	p=0.2247	Holm (x10) : p=1
	GLM RPE(M3) weight vs 0 : Proba Start	one sample Wilcoxon	p=0.2334	Holm (x10) : p=1
	GLM RPE(M3) weight vs 0 : Proba Mid	one sample Wilcoxon	**p=0.0049**	**Holm (x10) : p=0.0472**
	GLM RPE(M3) weight vs 0 : Proba Late	one sample Student t-test	**p=0.0047**	**Holm (x10) : p=0.0472**
	GLM RPE(M3) weight vs 0 : Proba Last	one sample Student t-test	**p=0.0140**	**Holm (x10) : p=0.0979**
A (bottom)	Success rate Trial Det_End vs Cplx_Start	Student unpaired t-test	**$p<10^{-5}$**	
	Success rate Trial Cplx Start vs Mid vs End	one way ANOVA	**p=0.0019**	
	Success rate Trial Cplx_End vs Proba_Start	Student unpaired t-test	**$p<10^{-5}$**	
	Success rate Trial Proba Start vs Mid vs Late vs Last	one way ANOVA	**p=0.0104**	
	Success rate Uturn Det_End vs Cplx_Start	Mann-Whitney U test (unpaired)	**$p<10^{-5}$**	
	Success rate Uturn Cplx Start vs Mid vs End	Kruskall-Wallis	p=0.6140	
	Success rate Uturn Cplx_End vs Proba_Start	Student unpaired t-test	**p=0.00003**	
	Success rate Uturn Proba Start vs Mid vs Late vs Last	Kruskall-Wallis	p=0.1553	
	Success rate Fwd Det_End vs Cplx_Start	Student unpaired t-test	**$p<10^{-5}$**	
	Success rate Fwd Cplx Start vs Mid vs End	Kruskall-Wallis	**p=0.0027**	
	Success rate Fwd Cplx_End vs Proba_Start	Mann-Whitney U test (unpaired)	**p=0.0004**	
	Success rate Fwd Proba Start vs Mid vs Late vs Last	Kruskall-Wallis	p=0.1553	
	Success rate pC Det_End vs Cplx_Start	Mann-Whitney U test (unpaired)	**$p<10^{-5}$**	
	Success rate pC-p100 Cplx Start vs Mid vs End	Kruskall-Wallis	**p=0.0070**	
	Success rate pC-p100 Cplx_End vs Proba_Start	Student unpaired t-test	**$p<10^{-5}$**	
	Success rate pB Det_End vs Cplx_Start	Student unpaired t-test	**$p<10^{-5}$**	
	Success rate pB-p50 Cplx Start vs Mid vs End	one way ANOVA	**p=0.0185**	
	Success rate pB-p50 Cplx_End vs Proba_Start	Student unpaired t-test	**$p<10^{-5}$**	
	Success rate pA Det_End vs Cplx_Start	Student unpaired t-test	**$p<10^{-5}$**	
	Success rate pA-p25 Cplx Start vs Mid vs End	one way ANOVA	**p=0.0029**	
	Success rate pA-p25 Cplx_End vs Proba_Start	Student unpaired t-test	**$p<10^{-5}$**	
B (left, middle)	Cplx post-reward dDA across sessions : Start (N=17) vs Mid (N=17) vs Last (N=17)	Kruskall-Wallis	p=0.9577	
	Cplx dDA : Start vs 0	one sample Student t-test	**p=0.0003**	**Holm (x3) : p=0.0003**
	Cplx dDA : Mid vs 0	one sample Wilcoxon	**p=0.00005**	**Holm (x3) : p=0.0001**
	Cplx dDA : Last vs 0	one sample Wilcoxon	**p=0.00002**	**Holm (x3) : p=0.00006**
B (left, bottom)	Cplx post-omission dDA across sessions : Start (N=17) vs Mid (N=17) vs Last (N=17)	one way ANOVA	p=0.1659	
	Cplx dDA : Start vs 0	one sample Student t-test	**p=0.00005**	**Holm (x3) : p=0.0002**
	Cplx dDA : Mid vs 0	one sample Student t-test	**p=0.0286**	**Holm (x3) : p=0.0286**
	Cplx dDA : Last vs 0	one sample Student t-test	**p=0.0018**	**Holm (x3) : p=0.0036**
B (right, middle)	Cplx across sessions : linear regression post-reward DA with success rate	Spearman correlation	p=0.2359 ; R2 = 0.073	
B (right, bottom)	Cplx across sessions : linear regression post-omission DA with sequence complexity	Pearson correlation	p=0.2037 ; R2 = 0.033	
C (left, middle)	Proba post-reward dDA across sessions : Start (N=12) vs Mid (N=12) vs Last (N=11) vs Last (N=10)	Kruskall-Wallis	**p=0.0092**	
	Proba dDA : Start vs 0	one sample Student t-test	p=0.8041	Holm (x4) : p=0.8041
	Proba dDA : Mid vs 0	one sample Wilcoxon	**p=0.0001**	**Holm (x4) : p=0.0039**
	Proba dDA : Late vs 0	one sample Wilcoxon	**p=0.0049**	**Holm (x4) : p=0.0146**
	Proba dDA : Last vs 0	one sample Wilcoxon	**p=0.0137**	**Holm (x4) : p=0.0125**
C (left, bottom)	Proba post-omission dDA across sessions : Start (N=12) vs Mid (N=12) vs Last (N=11) vs Last (N=10)	Kruskall-Wallis	**p=0.0651**	
	Proba dDA : Start vs 0	one sample Student t-test	p=0.9590	Holm (x4) : p=0.9590
	Proba dDA : Mid vs 0	one sample Wilcoxon	p=0.1294	Holm (x4) : p=0.2588
	Proba dDA : Late vs 0	one sample Student t-test	**p=0.0225**	**Holm (x4) : p=0.0676**
	Proba dDA : Last vs 0	one sample Student t-test	**p=0.0020**	**Holm (x4) : p=0.0078**
C (right, middle)	Proba across sessions : linear regression post-reward dDA with exploitation index	Spearman correlation	**$p<10^{-5}$; R2 = 0.1660**	
C (right, bottom)	Proba across sessions : linear regression post-omission dDA with #Success	Spearman correlation	**p=0.0040 ; R2 = 0.1423**	

Supp 1

Panel	Comparison	Test type	p-values	Corrections
A (left)	#Trials, male (N=23) vs female (N=26), S1-2 vs S4-5 vs Last2 (N=3 repeated measures)	mixed ANOVA (sex X session effect, with repeated measures on sessions)	Sex effect: p=0.7131 **Session effect: $p<10^{-5}$** Interaction effect: p=0.0561	
	#Trials in Det Last2, male (N=23) vs female (N=26)	Student unpaired t-test	p=0.3647	
A (center-left)	%Uturn in Det, male (N=23) vs female (N=26), S1-2 vs S4-5 vs Last2 (N=3 repeated measures)	mixed ANOVA (sex X session effect, with repeated measures on sessions)	Sex effect: p=0.5300 **Session effect: $p<10^{-5}$** Interaction effect: p=0.1597	
	%Uturn in Det Last2, male (N=23) vs female (N=26)	Student unpaired t-test	p=0.3469	
A (center-right)	Sequence cplx in Det, male (N=23) vs female (N=26), S1-2 vs S4-5 vs Last2 (N=3 repeated measures)	mixed ANOVA (sex X session effect, with repeated measures on sessions)	Sex effect: p=0.1672 **Session effect: $p<10^{-5}$** Interaction effect: p=0.1952	
	Sequence cplx in Det Last2, male (N=23) vs female (N=26)	Mann-Whitney U-test	**p=0.0346**	
A (right)	Circularity index in Det, male (N=23) vs female (N=26), S1-2 vs S4-5 vs Last2 (N=3 repeated measures)	mixed ANOVA (sex X session effect, with repeated measures on sessions)	Sex effect: p=0.2553 **Session effect: p=0.0022** Interaction effect: p=0.3185	
	Circularity index in Det Last2, male (N=23) vs female (N=26)	Mann-Whitney U-test	p=0.2255	
B (left)	%Success in Cplx, male (N=23) vs female (N=26), S1-2 vs S4-5 vs Last2 (N=3 repeated measures)	mixed ANOVA (sex X session effect, with repeated measures on sessions)	Sex effect: p=0.3353 **Session effect: $p<10^{-5}$** Interaction effect: p=0.0717	
	%Success in Cplx Last2, male (N=23) vs female (N=26)	Student unpaired t-test	p=0.5886	
B (center-left)	%Uturn in Cplx, male (N=23) vs female (N=26), S1-2 vs S4-5 vs Last2 (N=3 repeated measures)	mixed ANOVA (sex X session effect, with repeated measures on sessions)	Sex effect: p=0.5934 **Session effect: $p<10^{-5}$** **Interaction effect: p=0.0087**	
	%Uturn in Cplx Last2, male (N=23) vs female (N=26)	Student unpaired t-test	p=0.6816	
B (center-right)	Sequence cplx in Cplx, male (N=23) vs female (N=26), S1-2 vs S4-5 vs Last2 (N=3 repeated measures)	mixed ANOVA (sex X session effect, with repeated measures on sessions)	**Sex effect: p=0.0462** **Session effect: p=0.0001** Interaction effect: p=0.7944	
	Sequence cplx in Cplx Last2, male (N=23) vs female (N=26)	Mann-Whitney U-test	p=0.2662	
B (right)	Circularity index in Cplx, male (N=23) vs female (N=26), S1-2 vs S4-5 vs Last2 (N=3 repeated measures)	mixed ANOVA (sex X session effect, with repeated measures on sessions)	Sex effect: p=0.3757 **Session effect: p= $p<10^{-5}$** Interaction effect: p=0.7407	
	Circularity index in Cplx Last2, male (N=23) vs female (N=26)	Mann-Whitney U-test	p=0.6961	
C (left)	%Success in Proba, male (N=23) vs female (N=26), S1-2 vs S4-5 vs Last2 (N=3 repeated measures)	mixed ANOVA (sex X session effect, with repeated measures on sessions)	Sex effect: p=0.8112 **Session effect: $p<10^{-5}$** Interaction effect: p=0.9326	
	%Success in Proba Last2, male (N=23) vs female (N=26)	Student unpaired t-test	p=0.4590	
C (center-left)	%Uturn in Proba, male (N=23) vs female (N=26), S1-2 vs S4-5 vs Last2 (N=3 repeated measures)	mixed ANOVA (sex X session effect, with repeated measures on sessions)	Sex effect: p=0.8129 **Session effect: $p<10^{-5}$** Interaction effect: p=0.2954	
	%Uturn in Proba Last2, male (N=23) vs female (N=26)	Mann-Whitney U-test	p=0.4770	
C (center-right)	Sequence cplx in Proba, male (N=23) vs female (N=26), S1-2 vs S4-5 vs Last2 (N=3 repeated measures)	mixed ANOVA (sex X session effect, with repeated measures on sessions)	Sex effect: p=0.2919 **Session effect: $p<10^{-5}$** Interaction effect: p=0.6391	
	Sequence cplx in Proba Last2, male (N=23) vs female (N=26)	Mann-Whitney U-test	p=0.3312	
C (right)	Circularity index in Proba, male (N=23) vs female (N=26), S1-2 vs S4-5 vs Last2 (N=3 repeated measures)	mixed ANOVA (sex X session effect, with repeated measures on sessions)	Sex effect: p=0.8792 **Session effect: p= $p<10^{-5}$** Interaction effect: p=0.4908	
	Circularity index in Proba Last2, male (N=23) vs female (N=26)	Student unpaired t-test	p=0.6511	

Supp 2

Panel	Comparison	Test type	p-values	Corrections
B	Frequency distribution of 10-length chains for all mice End sessions, Det (n=2129 seq) vs Cplx (n=2838)	Kolmogorov-Smirnov (distribution)	**$p<10^{-5}$**	

Supp 3

Panel	Comparison	Test type	p-values	Corrections
A	#Trials across time point (N=49 mice, all paired)	one-way ANOVA (time point effect)	**$p<10^{-5}$**	
	Post-hoc, #Trials Det: First2 vs Last2 (N=49)	Wilcoxon (paired)	**$p<10^{-5}$**	**Holm (x5) : $p<10^{-5}$**
	Post-hoc, #Trials: Det Last2 vs Cplx First2 (N=49)	Student paired t-test	**$p<10^{-5}$**	**Holm (x5) : $p<10^{-5}$**
	Post-hoc, #Trials Cplx: First2 vs Last2 (N=49)	Student paired t-test	**$p<10^{-5}$**	**Holm (x5) : $p<10^{-5}$**
	Post-hoc, #Trials: Cplx Last2 vs Proba First2 (N=49)	Student paired t-test	p=0.5305	Holm (x5) : p=0.5305
	Post-hoc, #Trials Proba: First2 vs Last2 (N=49)	Student paired t-test	**p=0.0105**	**Holm (x5) : p=0.0210**
B (left)	Correlation between behavioural parameters in Det Last2 (N=49)	Pearson if normal, Spearman if not	See colour code in figure	Bonferroni (x66), see figure
B (center)	Correlation between behavioural parameters in Cplx First2 and Last2 (N=49)	Pearson if normal, Spearman if not	See colour code in figure	Bonferroni (x78), see figure
B (right)	Correlation between behavioural parameters in Proba First2 and Last2 (N=49)	Pearson if normal, Spearman if not	See colour code in figure	Bonferroni (x91), see figure

Supp 4

Panel	Comparison	Test type	p-values	Corrections
E	Det Last, Post-ICSS avg per mouse (N=14), Expected (task) vs Unexpected (restcage) (individual data from Fig2.E.)	Student paired t-test	**p=0.0019**	
F	Det Last, Post-ICSS avg per mouse (N=13), Expected (on-target) vs Unexpected (off-target) (individual data from Fig2.F.)	Student paired t-test	**$p<10^{-5}$**	
G	Det S1, Post-ICSS avg per mouse (N=9), task vs restcage	Wilcoxon (paired)	p=0.1641	
H	Det S2, Post-ICSS avg per mouse (N=9), on-target vs off-target	Student paired t-test	**p=0.0430**	

Supp 5

Panel	Comparison	Test type	p-values	Corrections
A	Restcage stimulation dF/F avg per mouse (N=14) : short (<3s) vs mid vs long (>5s)	one-way ANOVA	p=0.1508	
B	Restcage stimulation GLM : Intercept weight vs 0 (N=14)	one sample Student t-test	**$p<10^{-5}$**	**Holm (x2) : $p<10^{-5}$**
	Restcage stimulation GLM : T_inter_stim weight vs 0 (N=14)	one sample Student t-test	p=0.1526	Holm (x2) : p=0.1526
C	Det End post-reward dF/F avg per mouse (N=19) : Uturn vs Forward (individual data from Fig3.C.)	Student paired t-test	**p=0.0012**	
D	Det End post-reward dF/F avg per mouse (N=19) : pA vs pB vs pC	one-way ANOVA	p=0.6686	
E (top)	Cplx End post-reward dF/F avg per mouse (N=17) : Reward prev=rew vs prev=omi (individual data from Fig3.E.)	Student paired t-test	**p=0.0003**	
E (bottom)	Cplx End post-reward dF/F avg per mouse (N=17) : Omission prev=rew vs prev=omi (individual data from Fig3.E.)	Student paired t-test	**p=0.0357**	
F (top)	Cplx End post-reward dF/F avg per mouse (N=17) : Reward p100 vs p50 vs p25	one-way ANOVA	p=0.8132	
F (bottom)	Cplx End post-reward dF/F avg per mouse (N=17) : Omission pA vs pB vs pC	one-way ANOVA	p=0.3823	
G (top)	Cplx End post-reward dF/F avg per mouse (N=17) : Reward Uturn vs Fwd	Student paired t-test	p=0.1901	
G (bottom)	Cplx End post-reward dF/F avg per mouse (N=17) : Omission Uturn vs Fwd	Student paired t-test	p=0.3378	
H (top)	Proba End post-reward dF/F avg per mouse (N=10) : p100 vs p50 vs p25 (individual data from Fig3.G.)	one-way ANOVA	**p=0.0364**	
	Post-hoc, p100 vs p50	Student paired t-test	**p=0.0171**	**Holm (x3) : p=0.0282**
	Post-hoc, p50 vs p25	Student paired t-test	**p=0.0141**	**Holm (x3) : p=0.0282**
	Post-hoc, p100 vs p25	Student paired t-test	**p=0.0079**	**Holm (x3) : p=0.0237**
H (bottom)	Proba End post-omission dF/F avg per mouse (N=10) : p50 vs p25 (individual data from Fig3.G.)	Student paired t-test	**p=0.0173**	
I (top)	Proba End post-reward dF/F avg per mouse (N=10) : Reward prev=rew vs prev=omi	Student paired t-test	**p=0.0161**	
I (bottom)	Proba End post-reward dF/F avg per mouse (N=10) : Omission prev=rew vs prev=omi	Student paired t-test	p=0.9324	
J (top)	Proba End post-reward dF/F avg per mouse (N=10) : Reward Uturn vs Fwd	Student paired t-test	p=0.1628	
J (bottom)	Proba End post-reward dF/F avg per mouse (N=10) : Omission Uturn vs Fwd	Student paired t-test	p=0.0840	

Supp 6

Panel	Comparison	Test type	p-values	Corrections
B (top-left)	Det End RPE(M1) avg per mouse (N=19) : pA vs pB vs pC	one-way ANOVA	p=0.9222	
B (top-right)	Det End RPE(M1) avg per mouse (N=19) : Uturn vs Fwd	Wilcoxon	**p=0.0204**	
B (center-left)	Det End RPE(M2) avg per mouse (N=19) : pA vs pB vs pC	one-way ANOVA	p=0.8711	
B (center-right)	Det End RPE(M2) avg per mouse (N=19) : Uturn vs Fwd (Same as Fig 4.C.)	Wilcoxon	**p=0.00002**	
B (bottom-left)	Det End RPE(M3) avg per mouse (N=19) : pA vs pB vs pC	one-way ANOVA	p=0.9801	
B (bottom-right)	Det End RPE(M3) avg per mouse (N=19) : Uturn vs Fwd	Student paired t-test	p=0.4844	
C (top-left)	Cplx End RPE(M1) avg per mouse (N=17) : pA vs pB vs pC	Rew: one-way ANOVA Omi: one-way ANOVA	Rew: p=0.9843 Omi: p=0.9697	
C (top-middle)	Cplx End RPE(M1) avg per mouse (N=17) : Uturn vs Fwd	Rew: Student paired t-test Omi: Student paired t-test	**Rew: p=0.0223** **Omi: p=0.0018**	
C (top-right)	Cplx End RPE(M1) avg per mouse (N=17) : prev=rew vs prev=omi (Same as Fig 4.F.)	Rew: Student paired t-test Omi: Student paired t-test	**Rew: p<10e-5** **Omi: p<10e-5**	
C (center-left)	Cplx End RPE(M2) avg per mouse (N=17) : pA vs pB vs pC	Rew: one-way ANOVA Omi: one-way ANOVA	Rew: p=0.9125 Omi: p=0.9327	
C (center-middle)	Cplx End RPE(M2) avg per mouse (N=17) : Uturn vs Fwd	Rew: Student paired t-test Omi: Student paired t-test	**Rew: p=0.0004** **Omi: p=0.0002**	
C (center-right)	Cplx End RPE(M2) avg per mouse (N=17) : prev=rew vs prev=omi	Rew: Student paired t-test Omi: Student paired t-test	Rew: p=0.9737 **Omi: p=0.00002**	
C (bottom-left)	Cplx End RPE(M3) avg per mouse (N=17) : pA vs pB vs pC	Rew: one-way ANOVA Omi: one-way ANOVA	Rew: p=0.5187 Omi: p=0.4841	
C (bottom-middle)	Cplx End RPE(M3) avg per mouse (N=17) : Uturn vs Fwd	Rew: Student paired t-test Omi: Student paired t-test	Rew: p=0.2845 Omi: p=0.0612	
C (bottom-right)	Cplx End RPE(M3) avg per mouse (N=17) : prev=rew vs prev=omi	Rew: Student paired t-test Omi: Student paired t-test	**Rew: p=0.0211** Omi: p=0.7743	
D (top-left)	Proba End RPE(M1) avg per mouse (N=10) : p100 vs p50 vs p25 for rewards, p50 vs p25 for omissions	Rew: one-way ANOVA Omi: Student paired t-test	Rew: p=0.8881 Omi: p=0.3362	
D (top-middle)	Proba End RPE(M1) avg per mouse (N=10) : Uturn vs Fwd	Rew: Student paired t-test Omi: Student paired t-test	**Rew: p=0.0085** Omi: p=0.1934	
D (top-right)	Proba End RPE(M1) avg per mouse (N=10) : prev=rew vs prev=omi	Rew: Student paired t-test Omi: Student paired t-test	**Rew: p<10e-5** **Omi: p=0.00005**	
D (center-left)	Proba End RPE(M2) avg per mouse (N=10) : p100 vs p50 vs p25 for rewards, p50 vs p25 for omissions	Rew: one-way ANOVA Omi: Student paired t-test	Rew: p=0.1195 **Omi: p=0.0042**	
D (center-middle)	Proba End RPE(M2) avg per mouse (N=10) : Uturn vs Fwd	Rew: Student paired t-test Omi: Student paired t-test	**Rew: p=0.00005** **Omi: p=0.00006**	
D (center-right)	Proba End RPE(M2) avg per mouse (N=10) : prev=rew vs prev=omi	Rew: Student paired t-test Omi: Student paired t-test	**Rew: p=0.0004** **Omi: p=0.0005**	
D (bottom-left)	Proba End RPE(M3) avg per mouse (N=10) : p100 vs p50 vs p25 for rewards, p50 vs p25 for omissions (Same as Fig 4.I.)	Rew: Kruskall-Wallis Omi: Student paired t-test	**Rew: p<10e-5** **Omi: p=0.0020**	
D (bottom-middle)	Proba End RPE(M3) avg per mouse (N=10) : Uturn vs Fwd	Rew: Student paired t-test Omi: Wilcoxon	**Rew: p=0.00006** **Omi: p=0.0020**	
D (bottom-right)	Proba End RPE(M3) avg per mouse (N=10) : prev=rew vs prev=omi	Rew: Student paired t-test Omi: Student paired t-test	**Rew: p<10e-5** Omi: p=0.6590	

Supp 7

Panel	Comparison	Test type	p-values	Corrections
A (top)	GLM RPE(M1) weight across Det	Kruskall-Wallis	p=0.6029	
	GLM RPE(M1) weight : End Det vs Start Cplx	Student unpaired t-test	**p=0.0001**	
	GLM RPE(M1) weight across Cplx	one-way ANOVA	p=0.9001	
	GLM RPE(M1) weight : End Cplx vs Start Proba	Student unpaired t-test	p=0.4669	
	GLM RPE(M1) weight across Proba	one-way ANOVA	p=0.0031	
	GLM RPE(M1) weight vs 0 : Det Mid	one sample Student t-test	p=0.4476	Holm (x10) : p=1
	GLM RPE(M1) weight vs 0 : Det Late	one sample Student t-test	p=0.6257	Holm (x10) : p=1
	GLM RPE(M1) weight vs 0 : Det Last	one sample Student t-test	p=0.4849	Holm (x10) : p=1
	GLM RPE(M1) weight vs 0 : Cplx Start	one sample Student t-test	**p=0.00003**	Holm (x10) : p=**0.0003**
	GLM RPE(M1) weight vs 0 : Cplx Mid	one sample Student t-test	**p=0.0004**	**Holm (x10) : p=0.0034**
	GLM RPE(M1) weight vs 0 : Cplx Last	one sample Student t-test	**p=0.0006**	**Holm (x10) : p=0.0047**
	GLM RPE(M1) weight vs 0 : Proba Start	one sample Student t-test	**p=0.0211**	Holm (x10) : p=0.1476
	GLM RPE(M1) weight vs 0 : Proba Mid	one sample Wilcoxon	p=0.4277	Holm (x10) : p=1
	GLM RPE(M1) weight vs 0 : Proba Late	one sample Student t-test	p=0.3216	Holm (x10) : p=1
	GLM RPE(M1) weight vs 0 : Proba Last	one sample Student t-test	p=0.3218	Holm (x10) : p=1
	GLM RPE(M2) weight across Det	Kruskall-Wallis	**p=0.0117**	
	GLM RPE(M2) weight : End Det vs Start Cplx	Mann-Whitney U test (unpaired)	p=0.1061	
	GLM RPE(M2) weight across Cplx	Kruskall-Wallis	p=0.6152	
	GLM RPE(M2) weight : End Cplx vs Start Proba	Mann-Whitney U test (unpaired)	**p=0.0096**	
	GLM RPE(M2) weight across Proba	one-way ANOVA	p=0.3649	
	GLM RPE(M2) weight vs 0 : Det Mid	one sample Student t-test	**p=0.0295**	Holm (x10) : p=0.2356
	GLM RPE(M2) weight vs 0 : Det Late	one sample Student t-test	**p=0.0006**	**Holm (x10) : p=0.0061**
	GLM RPE(M2) weight vs 0 : Det Last	one sample Student t-test	p=0.9577	Holm (x10) : p=1
	GLM RPE(M2) weight vs 0 : Cplx Start	one sample Wilcoxon	**p=0.0174**	Holm (x10) : p=0.1569
	GLM RPE(M2) weight vs 0 : Cplx Mid	one sample Wilcoxon	p=0.1324	Holm (x10) : p=0.6619
	GLM RPE(M2) weight vs 0 : Cplx Last	one sample Wilcoxon	p=0.0569	Holm (x10) : p=0.3982
	GLM RPE(M2) weight vs 0 : Proba Start	one sample Student t-test	p=0.0681	Holm (x10) : p=0.4085
	GLM RPE(M2) weight vs 0 : Proba Mid	one sample Student t-test	p=0.2847	Holm (x10) : p=1
	GLM RPE(M2) weight vs 0 : Proba Late	one sample Student t-test	p=0.8733	Holm (x10) : p=1
	GLM RPE(M2) weight vs 0 : Proba Last	one sample Student t-test	p=0.9266	Holm (x10) : p=1
	GLM RPE(M3) weight across Det	one-way ANOVA	p=0.3810	
	GLM RPE(M3) weight : End Det vs Start Cplx	Student unpaired t-test	p=0.6093	
	GLM RPE(M3) weight across Cplx	Kruskall-Wallis	p=0.4623	
	GLM RPE(M3) weight : End Cplx vs Start Proba	Mann-Whitney U test (unpaired)	p=0.25883	
	GLM RPE(M3) weight across Proba	Kruskall-Wallis	**p=0.0761**	
	GLM RPE(M3) weight vs 0 : Det Mid	one sample Student t-test	p=0.3899	Holm (x10) : p=1
	GLM RPE(M3) weight vs 0 : Det Late	one sample Student t-test	p=0.9362	Holm (x10) : p=1
	GLM RPE(M3) weight vs 0 : Det Last	one sample Student t-test	p=0.5057	Holm (x10) : p=1
	GLM RPE(M3) weight vs 0 : Cplx Start	one sample Student t-test	p=0.1089	Holm (x10) : p=0.7621
	GLM RPE(M3) weight vs 0 : Cplx Mid	one sample Wilcoxon	p=0.4529	Holm (x10) : p=1
	GLM RPE(M3) weight vs 0 : Cplx Last	one sample Wilcoxon	p=0.9632	Holm (x10) : p=1
	GLM RPE(M3) weight vs 0 : Proba Start	one sample Wilcoxon	p=0.1099	Holm (x10) : p=0.7621
	GLM RPE(M3) weight vs 0 : Proba Mid	one sample Wilcoxon	**p=0.0015**	**Holm (x10) : p=0.0146**
	GLM RPE(M3) weight vs 0 : Proba Late	one sample Student t-test	**p=0.0050**	**Holm (x10) : p=0.0449**
	GLM RPE(M3) weight vs 0 : Proba Last	one sample Student t-test	**p=0.0124**	**Holm (x10) : p=0.0988**
B (left)	Det %Uturns (N=19) : S1-2 vs S4-5 vs Last2	one-way ANOVA	**$p<10^{-5}$**	
B (right)	Det seq. cplx (N=19) : S1-2 vs S4-5 vs Last2	one-way ANOVA	**$p<10^{-5}$**	
C	Det dDA across sessions : Start (N=9) vs Mid (N=9) vs Late (N=14) vs Last (N=19)	Kruskall-Wallis	p=0.3857	
	Det dDA : Start vs 0	one sample Student t-test	p=0.3472	Holm (x4) : p=0.3472
	Det dDA : Mid vs 0	one sample Student t-test	p=0.0990	Holm (x4) : p=0.1980
	Det dDA : Late vs 0	one sample Student t-test	**p=0.0353**	Holm (x4) : p=0.1058
	Det dDA : Last vs 0	one sample Wilcoxon	**p=0.0033**	**Holm (x4) : p=0.0134**
D	Det across sessions : linear regression post-reward DA with turn rate	Spearman correlation	p=0.9643 ; R2 = 0.0025	
E (center-left)	Cplx %Success (N=18) : S1-2 vs S4-5 vs Last2	one-way ANOVA	**$p<10^{-5}$**	
E (center)	Cplx %Uturns (N=18) : S1-2 vs S4-5 vs Last2	one-way ANOVA	**$p<10^{-5}$**	
E (center-right)	Cplx seq. cplx (N=18) : S1-2 vs S4-5 vs Last2	one-way ANOVA	**$p<10^{-5}$**	
E (right)	Cplx circularity index (N=18) : S1-2 vs S4-5 vs Last2	one-way ANOVA	**$p<10^{-5}$**	
F (center-left)	Proba %Success (N=14) : S1-2 vs S4-5 vs Last2	one-way ANOVA	**$p<10^{-5}$**	
F (center)	Proba %Uturns (N=14) : S1-2 vs S4-5 vs Last2	one-way ANOVA	**$p<10^{-5}$**	
F (center-right)	Proba seq. cplx (N=14) : S1-2 vs S4-5 vs Last2	one-way ANOVA	**$p<10^{-5}$**	
F (right)	Proba exploitation index (N=14) : S1-2 vs S4-5 vs Last2	one-way ANOVA	**$p<10^{-5}$**	

Supp 8

Panel	Comparison	Test type	p-values	Corrections
A (left)	Cplx Start: linear regression between DA reward peak and length of reward chains (n=1812)	Spearman correlation	**$p<10^{-5}$; R2 = 0.009**	
	Cplx Start: linear regression between DA reward peak and length of omission chains (n=385)	Spearman correlation	**p=0.00005 ; R2 = 0.083**	
A (right)	Cplx Start: linear regression between DA omission dip and length of reward chains (n=395)	Spearman correlation	**p=0.00002 ; R2 = 0.030**	
	Cplx Start: linear regression between DA omission dip and length of omission chains (n=816)	Spearman correlation	**$p<10^{-5}$; R2 = 0.088**	
B (left)	Cplx End: linear regression between DA reward peak and length of reward chains (n=2815)	Spearman correlation	p=0.5887 ; R2 = 0.001	
	Cplx End: linear regression between DA reward peak and length of omission chains (n=604)	Spearman correlation	**p=0.00001 ; R2 = 0.015**	
B (right)	Cplx End: linear regression between DA omission dip and length of reward chains (n=612)	Spearman correlation	**p=0.0002 ; R2 = 0.019**	
	Cplx End: linear regression between DA omission dip and length of omission chains (n=666)	Spearman correlation	**p=0.0028 ; R2 = 0.008**	
C (left)	Cplx Start: Success rate depending on previous choice : forward vs uturn (N=17)	Student paired t-test	**p<10-5**	
C (right)	Cplx End: Success rate depending on previous choice : forward vs uturn (N=17)	Student paired t-test	**p=0.0005**	
D (left)	Cplx Start: Uturn rate depending on previous outcome : reward vs omission (N=17)	Student paired t-test	p=0.1443	
D (right)	Cplx End: Uturn rate depending on previous outcome : reward vs omission (N=17)	Student paired t-test	p=0.0577	
E (bottom, left)	Outcome chains length for all trials Cplx sessions : Rew_First2 (n=3472) vs Omi_First2 (n=2247) vs Rew_Last2 (n=6947) vs Omi_Last2 (n=2463) (N=49)	Kolmogorov-Smirnov (distribution)	**First2: Rew vs Omi: $p<10^{-5}$** **Last2: Rew vs Omi: $p<10^{-5}$** **Rew: First2 vs Last2: $p<10^{-5}$** **Omi: First2 vs Last2: $p<10^{-5}$**	**Holm (x4) : all $p<10^{-5}$**
E (bottom, right)	Outcome chains length for all mice Cplx sessions : Rew_First2 vs Omi_First2 vs Rew_Last2 vs Omi_Last2 (N=49)	one way ANOVA	**$p<10^{-5}$**	
	Post-hoc Rew_First2 vs Omi_First2:	Wilcoxon test	**p=0.0230**	**Holm (x4) : p=0.0230**
	Post-hoc Rew_Last2 vs Omi_Last2:	Wilcoxon test	**$p<10^{-5}$**	**Holm (x4) : $p<10^{-5}$**
	Post-hoc Rew_First2 vs Rew_Last2:	Wilcoxon test	**$p<10^{-5}$**	**Holm (x4) : $p<10^{-5}$**
	Post-hoc Omi_First2 vs Omi_Last2:	Wilcoxon test	**$p<10^{-5}$**	**Holm (x4) : $p<10^{-5}$**
F (left)	Uturn chains length for all trials Cplx sessions : Fwd_First2 (n=4235) vs Uturn_First2 (n=1386) vs Fwd_Last2 (n=6137) vs Uturn_Last2 (n=3175) (N=49)	Kolmogorov-Smirnov (distribution)	**First2: Fwd vs Uturn: $p<10^{-5}$** **Last2: Fwd vs Uturn: $p<10^{-5}$** **Fwd: First2 vs Last2: $p<10^{-5}$** Uturn: First2 vs Last2: p=0.9501	**Holm (x4) : $p<10^{-5}$** **Holm (x4) : $p<10^{-5}$** **Holm (x4) : $p<10^{-5}$** Holm (x4) : p=0.9501
F (right)	Uturn chains length for all mice Cplx sessions : Fwd_First2 vs Uturn_First2 vs Fwd_Last2 vs Uturn_Last2 (N=49)	one way ANOVA	**$p<10^{-5}$**	
	Post-hoc Fwd_First2 vs Uturn_First2:	Wilcoxon test	**$p<10^{-5}$**	**Holm (x4) : $p<10^{-5}$**
	Post-hoc Fwd_Last2 vs Uturn_Last2:	Wilcoxon test	**$p<10^{-5}$**	**Holm (x4) : $p<10^{-5}$**
	Post-hoc Fwd_First2 vs Fwd_Last2:	Wilcoxon test	**$p<10^{-5}$**	**Holm (x4) : $p<10^{-5}$**
	Post-hoc Uturn_First2 vs Uturn_Last2:	Wilcoxon test	p=0.6860	Holm (x4) : p=0.6860

References Supplementary Materials

1. H. Khabou, M. Garita-Hernandez, A. Chaffiol, S. Reichman, C. Jaillard, E. Brazhnikova, S. Bertin, V. Forster, M. Desrosiers, C. Winckler, O. Goureau, S. Picaud, J. Duebel, J.-A. Sahel, D. Dalkara, Noninvasive gene delivery to foveal cones for vision restoration. *JCI insight* **3**, D358 (2018).

2. F. Sun, J. Zeng, M. Jing, J. Zhou, J. Feng, S. F. Owen, Y. Luo, F. Li, H. Wang, T. Yamaguchi, Z. Yong, Y. Gao, W. Peng, L. Wang, S. Zhang, J. Du, D. Lin, M. Xu, A. C. Kreitzer, G. Cui, Y. Li, A Genetically Encoded Fluorescent Sensor Enables Rapid and Specific Detection of Dopamine in Flies, Fish, and Mice. *Cell* **174**, 481-496.e19 (2018).

3. F. Sun, J. Zhou, B. Dai, T. Qian, J. Zeng, X. Li, Y. Zhuo, Y. Zhang, Y. Wang, C. Qian, K. Tan, J. Feng, H. Dong, D. Lin, G. Cui, Y. Li, Next-generation GRAB sensors for monitoring dopaminergic activity in vivo. *Nat. methods* **17**, 1156–1166 (2020).

4. V. W. Choi, A. Asokan, R. A. Haberman, R. J. Samulski, Production of recombinant adeno-associated viral vectors. *Current protocols in human genetics* **Chapter 12**, Unit 12.9-12.9.21 (2007).

5. C. Aurnhammer, M. Haase, N. Muether, M. Hausl, C. Rauschhuber, I. Huber, H. Nitschko, U. Busch, A. Sing, A. Ehrhardt, A. Baiker, Universal real-time PCR for the detection and quantification of adeno-associated virus serotype 2-derived inverted terminal repeat sequences. *Human gene therapy methods* **23**, 18–28 (2012).

6. M. Belkaid, E. Bousseyrol, R. D. Cuttoli, M. Dongelmans, E. K. Duranté, T. A. Yahia, S. Didienne, B. Hanesse, M. Come, A. Mourot, J. Naudé, O. Sigaud, P. Faure, Mice adaptively generate choice variability in a deterministic task. *Communications Biology* **3**, 1–9 (2020).

7. J. Naudé, S. Tolu, M. Dongelmans, N. Torquet, S. Valverde, G. Rodriguez, S. Pons, U. Maskos, A. Mourot, F. Marti, P. Faure, Nicotinic receptors in the ventral tegmental area promote uncertainty-seeking. *Nature Neuroscience* **19**, 471–478 (2016).

8. E. Bousseyrol, S. Didienne, S. Takillah, C. Prevost-Solié, M. Come, T. A. Yahia, S. Mondoloni, E. Vicq, L. Tricoire, A. Mourot, J. Naudé, P. Faure, Dopaminergic and prefrontal dynamics co-determine mouse decisions in a spatial gambling task. *Cell Rep.* **42**, 112523 (2023).

9. M. Dongelmans, R. D. Cuttoli, C. Nguyen, M. Come, E. K. Duranté, D. Lemoine, R. Brito, T. A. Yahia, S. Mondoloni, S. Didienne, E. Bousseyrol, B. Hannesse, L. M. Reynolds, N. Torquet, D. Dalkara, F. Marti, A. Mourot, J. Naudé, P. Faure, Chronic nicotine increases midbrain dopamine neuron activity and biases individual strategies towards reduced exploration in mice. *Nat Commun* **12**, 6945 (2021).

10. A. Lempel, J. Ziv, On the Complexity of Finite Sequences. *IEEE Trans. Information Theory* **22**, 75–81 (1976).