

# Dopamine signaling and the distal reward problem

Douglas A. Nitz, William J. Kargo and Jason Fleischer

The Neurosciences Institute, John J. Hopkins Dr, San Diego, California, USA

Correspondence to Douglas A. Nitz, PhD, The Neurosciences Institute, 10640 John J. Hopkins Dr, San Diego, California, 92121, USA  
Tel: +1 858 626 2115; e-mail: nitz@nsi.edu

Received 21 July 2007; accepted 17 August 2007

Actions and their associated consequences, such as reward attainment, are often temporally distant. Animals nevertheless learn such associations thereby solving the 'distal reward' problem. We sought to determine whether dopamine signaling plays a role in such learning. Wild-type and dopamine type 1 receptor knockout mice executed three left/right choices leading to one of eight differentially rewarded goal sites. Compared with wild-type mice, knockouts exhibited selective impairments in decision making at

choice points distal, but not proximal, to goal sites. We conclude that dopamine's role in reinforcement learning depends on the temporal relationship of actions to reward and that dopamine signaling through D1 receptors constitutes a component of those brain mechanisms responsible for solving the distal reward problem. *NeuroReport* 18:1833–1836 © 2007 Wolters Kluwer Health | Lippincott Williams & Wilkins.

**Keywords:** decision making, distal reward, dopamine, knockout, reward

## Introduction

Defining the role of dopamine (DA) in learning and decision-making is critical to understanding attention deficit disorder, schizophrenia, and addiction where alterations in DA type 1 (D1) receptor function have been identified [1–3]. One method to identify dysfunction attributable to D1 receptors is to examine behavior in D1 receptor knockout (D1KO) mice.

DA plays a significant role in reinforcement learning [4,5]. Impairments are observed following lesion of DA neurons [6] and in Parkinson's disease [7] which is characterized by degeneration of DA neurons. Nevertheless, D1KO mice do not exhibit deficits in associative learning [8] and readily develop goal-directed behaviors [9]. This paradox might be explained by adaptations in brain development, but showing DA's role in reinforcement learning could require tasks of greater complexity. Suggesting this is: (i) the complex nature of DA signaling in prefrontal cortex [10] and (2) the finding that D1KO's show impairments in the Morris water tank navigational task [11].

We sought to detect subtle differences in behavioral ability between D1KO and wild-type (WT) mice that might resolve the aforementioned paradox and, more generally, to provide clues as to the role of DA in brain function. We considered the possibility that DA plays a role in action–consequence association, and the importance of the temporal distance between actions and consequences. That is, does DA signaling constitute part of the mechanism by which the brain solves the 'distal reward' problem as described by Hull [12].

## Methods

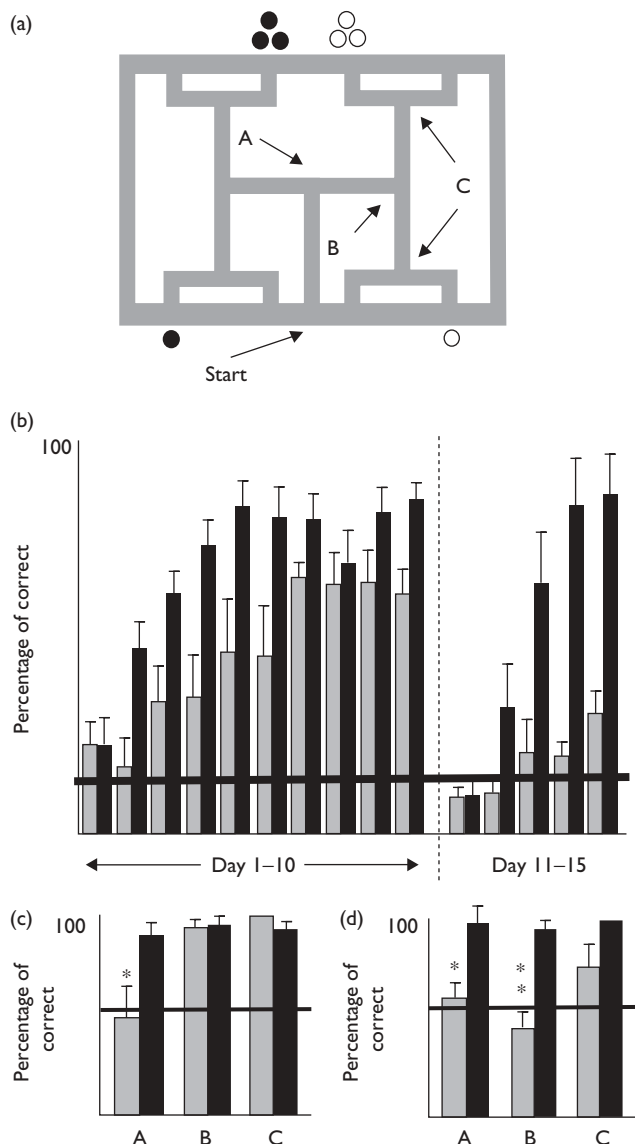
Groups were composed of six adult male mice (25–30 g, 210–270 days old). Heterozygous D1 mice (Jackson Laboratory, Bar Harbor, Maine) were bred to obtain homozygous D1KO

and WT littermates. Mice were housed 2–4 per cage, maintained on a 12:12 light–dark schedule, and genotyped using genomic tail DNA.

Mice were food-restricted to 85% of their free-feeding weights and familiarized with 10-mg chocolate pellets (Research Diets, New Brunswick, New Jersey, USA). Mice were introduced to the maze where pellets were randomly scattered to induce foraging. Mice were then trained to make traversals from the start line to any of eight goal sites where a small (1 pellet) reward was given (Fig. 1a). After a week of such training, reward was made available at only two sites. At the high reward site (HR), three pellets were given. A small reward (SR, 1 pellet) was given at another site. No reward was given at the remaining six sites. HR and SR goal sites remained the same for 10 days. Mice made 15 traversals/day. The maze was washed with quatricide solution before each set of traversals. Mice were not allowed to reverse course. This rule was enforced by placing a block behind the animal as it moved past a turn. On day 11, the HR–SR goal sites were switched to new locations and 5 days of additional experiments were completed.

## Analyses

The percentage of trials for which mice obtained either the SR or HR was determined as was the number of correct L–R choices at turns A, B, and C. As both reward sites were on the same side of the maze, incorrect choices at turn A precluded assessment of turn B and C behavior. An incorrect choice at B precluded assessment of turn C behavior. Scoring behavior at B was complicated by the fact that three mice (1 WT and 2 D1KO) moved primarily to the SR site. Whether mice moved primarily to the SR or HR site, choice was always strongly biased toward one or the other site. As such, turn B choices were considered correct when made toward the favored goal site.



**Fig. 1** Task schematics and choice accuracy. (a) The 'triple-T' maze ( $65 \times 80$  cm). On days 1–10, HR and SR (three black and one black filled circles, respectively) were available at two goal sites. On day 11, HR and SR sites were switched to the sites indicated by open circles. (b) Percentage of trials ( $\pm$ SD) for which WT (black) and D1KO (grey) mice reached either the HR or SR goal sites. Horizontal bar represents % correct expected by chance. (c) Turn choice accuracy on day 10. (d) Turn choice accuracy on day 15.

For each turn, we obtained a measure of the trial-to-trial probability of making a correct choice [13]. The analysis consists of a state-space model where an observable Bernoulli variable, a correct or incorrect choice, is used to estimate a hidden Gaussian variable, the probability of making the correct choice. The analysis returns a maximum-likelihood estimate of the learning curve. Group performance (Fig. 2) is the mean of individual learning curves.

## Results

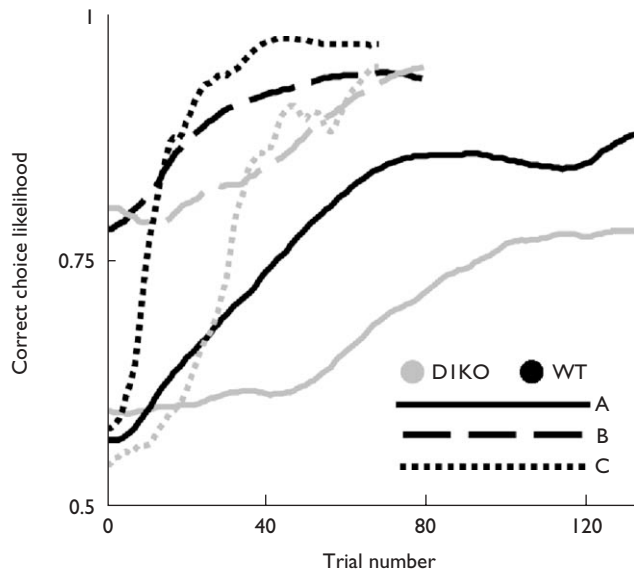
Both WT and D1KO mice executed sequences of L-R choices yielding reward. By day 10, the percentage of traversals to HR or SR sites was well above chance for all animals (one sample *t*-test,  $P < 0.01$  for both groups, Fig. 1b). An

analysis of variance with training day and mouse type as factors, however, yielded significant F values for training day ( $P < 0.001$ ), mouse type ( $P < 0.03$ ), and their interaction ( $P < 0.01$ ). Overall, D1KO mice traversed rewarded paths significantly less often indicating a deficiency in L-R choice behavior. Post hoc *t*-tests showed that this deficiency was apparent as early as the second day of training and persisted to day 10 of training when, for WT mice, the percentage of rewarded trials reached an asymptote at 85.5%.

We next asked whether choice behavior on day 10 was equally impaired at all three turns (A, B, and C). Turn C choices occur with the greatest proximity in space to goal sites (Fig. 1a). Turn A choices are most distal while turn B choices lie intermediate. Figure 1c depicts percent correct choice behavior at each turn on day 10. Although the percentage of correct choices at turns B and C were virtually the same for the two groups, D1KO mice exhibited an impairment at turn A (*t*-test,  $P < 0.02$ ). Thus, the impaired performance of D1KO mice in executing turn sequences leading to reward was largely attributable to deficiencies in L-R choice behavior at the turn (A) most distal to the goal site.

Performance of WT and D1KO mice also differed on days 11–15 when the SR and HR sites were changed (Fig. 1b, analysis of variance,  $P < 0.01$ ). On day 11, both WT and D1KO reached reward sites less often than expected by chance (one sample *t*-test,  $P < 0.01$  for each group). This was attributable to persistence in executing sequences leading to the former reward sites. By day 15, WT mice reached rewarded sites at rates similar to day 10. On day 15, however, D1KO mice still reached reward sites at rates close to chance. Again, choice behavior at turns A, B, and C suggested that proximity of choice points to goal sites is related to the ability of D1KO mice to learn action–outcome contingencies. On day 15, the percentage of correct choices at A and B remained close to chance levels for D1KO mice and were significantly different than those for WT mice (*t*-test,  $P < 0.02$  and  $P < 0.01$ , respectively). Turn C choice accuracy approached that for WTs and was not significantly different.

The foregoing data suggest that D1KO impairments in executing correct L-R choice sequences are attributable to deficits in learning action–outcome contingencies when actions and their outcomes (reward obtainment or lack thereof) are widely separated in time. One might then expect WT mice to learn action–outcome contingencies more efficiently when action and outcome are closely spaced. To test this, we applied an algorithm (see methods) that estimates the probability of making correct choices on the basis of the empirically observed series of correct and incorrect choices. Depicted in Fig. 2 is the maximum likelihood estimate of the probability that mice will make a correct choice as a function of trial number. The steepness of each curve is related to the rate at which learning occurred. The learning rate for turn A lags behind that for turn C despite the increased number of learning opportunities (animals had the opportunity to make a choice at turn A on every trial). The difference in learning curves can be quantified by examining the overlap between their 95% confidence intervals. Turn A and C confidence intervals (not shown on the figure for clarity) overlap during the first trials. They, however, no longer overlap for WT mice starting at trial 36, and overlap again at trial 52 as turn A behavior continues to improve. Thus, learning of action–outcome



**Fig. 2** Learning rates at choice points A, B, and C. For days 1–10 of the experiment, each line depicts the maximum likelihood estimate of the animals' probability of making a correct choice (WT-black, D1KO-grey). WT, wild-type.

contingencies is closely related to the distance between action and outcome.

## Discussion

The rate at which both WT and D1KO mice learned action–outcome contingencies depended on the distance between the site of action (i.e. the A, B, or C LR choice point) and the site where the outcome of those actions could be realized (goal sites where reward was, or was not, received, see Fig. 2). This finding indicates that solving the distal reward problem (i.e. associating actions and outcomes separated in time) becomes more difficult as the temporal distance between action and outcome increases. We assume that the critical difference was the temporal distance between choice points and the goal site. Nevertheless, it is conceivable that either the spatial distance or the ordering of choices is the determinant of learning rate.

Behavioral deficits exhibited by D1KO mice depended on the distance between maze positions where choices were made and goal sites where the consequences of those choices could be realized (Fig. 1c). Both WT and D1KO readily learned action–reward associations when the distance between the point of execution and obtainment of reward was minimal. D1KO mice exhibited impaired choice behavior at maze positions distant from reward obtainment. The severe impairment in D1KOs in adapting to alterations in action–reward contingencies (Fig. 1b, days 11–15) suggests that D1 receptor activation plays a role not only in the development, but also the extinction of action–reward associations. Together, the findings suggest that DA signaling through the type I receptor plays a role in binding together actions and their consequences which are separated in time. The time interval-dependent nature of learning deficits in D1KO mice likely explains earlier data indicating that D1KO mice can learn simple associations, but exhibit impairments in developing complex goal-directed action sequences. Associations between actions

and rewards proximal in time of occurrence could depend on other neuromodulatory systems or are possibly learned through spike timing-dependent plasticity if representations for actions and rewards are temporally coexistent.

The set of possible mechanisms by which activation of D1 receptors could contribute to associative learning is likely constrained by the dynamics of DA signaling. Recent evidence indicates that DA release is tonically enhanced throughout the performance of navigational tasks [14]. As DA modulates the signaling properties of prefrontal cortical and caudate neurons [15,16], D1KO deficits could reflect an inability to maintain, in these regions, representations of actions and rewards. In this respect, it is notable that prefrontal cortical activity in rodents in part reflects the history of behavior [17]. It follows, then, that behaviors further removed from the time of reward would be the most vulnerable.

A second possibility concerns deficits in phasic DA reward signaling in D1KO mice. Work in rats shows that DA neurons are phasically activated at the time of reward delivery [18]. In WT, but not D1KO mice, some nucleus accumbens neurons exhibit activity peaks near the time of reward [9]. Obviously, deficits in phasic reward signaling will have implications for almost any learning model. In particular, however, 'temporal difference' models seek to explain associative learning of temporally distant actions and rewards [19]. In such models, the probability of producing activity associated with correct decisions is increased as a function of phasic reward signaling which itself is highly sensitive to reward magnitude [20]. Presumably, activity driving correct motor decisions arises through changes in synaptic efficacy. Work with D1 receptor antagonists suggests a role for D1 receptors in this process [21]. DA-dependent changes in synaptic efficacy most likely occur in prefrontal cortex and striatum, where innervation by DA neurons is present and actions, stimuli, and rewards are all reflected in neuronal activity [22,23]. With respect to the present findings, it is notable that some temporal difference models predict faster learning of action–reward contingencies when actions and rewards are temporally proximal [24]. Thus, the known deficits in phasic reward signaling in D1KO mice could result in learning deficits dependent on the temporal distance between actions and their consequences. Recent theoretical work examining the relationship between spike timing-dependent plasticity and DA signaling is consistent with this assertion and may identify the mechanisms underlying the presently observed behavioral results [25].

## Conclusion

The rate at which an action–reward contingency is learned is related to the time between that action and reward delivery. D1KO mice exhibit deficits in learning of action–reward contingencies that depend on the temporal proximity of action and reward. These findings suggest that DA signaling through D1 receptors impacts brain processes linking actions and their temporally distal consequences.

## Acknowledgements

The authors thank Kara Papaefthimiou and Glen Davis for help in conducting experiments and organizing analyses. This work was supported by the Neurosciences Research

Foundation and the G. Harold and Leila Y. Mathers Charitable Foundation. W.J. Kargo is the Clayson Fellow in Motor Control.

## References

1. Okubo Y, Suhara T, Suzuki K, Kobayashi K, Inoue O, Terasaki O, *et al.* Decreased prefrontal dopamine D1 receptors in schizophrenia revealed by PET. *Nature* 1997; **385**:634–636.
2. Misener VL, Luca P, Azeke O, Crosbie J, Waldman I, Tannock R, *et al.* Linkage of the dopamine receptor D1 gene to attention-deficit/hyperactivity disorder. *Mol Psychiatry* 2004; **9**:500–509.
3. Bobb AJ, Addington AM, Sidransky E, Gornick MC, Lerch JP, Greenstein DK, *et al.* Support for association between ADHD and two candidate genes: NET1 and DRD1. *Am J Med Genet B Neuropsychiatr Genet* 2005; **134**:67–72.
4. Redish AD. Addiction as a computational process gone awry. *Science* 2004; **306**:1944–1947.
5. Clark L, Cools R, Robbins TW. The neuropsychology of ventral prefrontal cortex: decision-making and reversal learning. *Brain Cogn* 2004; **55**: 41–53.
6. Parkinson JA, Dalley JW, Cardinal RN, Bamford A, Fehnert B, Lachenal G, *et al.* Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive Pavlovian approach behaviour: implications for mesoaccumbens dopamine function. *Behav Brain Res* 2002; **137**:149–163.
7. Frank MJ, Seeberger LC, O'reilly RC. By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* 2004; **306**:1940–1943.
8. Smith DR, Striplin CD, Geller AM, Mailman RB, Drago J, Lawler CP, Gallagher M. Behavioural assessment of mice lacking D1A dopamine receptors. *Neuroscience* 1998; **86**:135–146.
9. Tran AH, Tamura R, Uwano T, Kobayashi T, Katsuki M, Ono T. Dopamine D1 receptors involved in locomotor activity and accumbens neural responses to prediction of reward associated with place. *Proc Natl Acad Sci U S A* 2005; **102**:2117–2122.
10. Seamans JK, Yang CR. The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Prog Neurobiol* 2004; **74**:1–58.
11. El-Ghundi M, Fletcher PJ, Drago J, Sibley DR, O'Dowd BF, George SR. Spatial learning deficit in dopamine D(1) receptor knockout mice. *Eur J Pharmacol* 1999; **383**:95–106.
12. Hull CL. *Principles of behavior: an introduction to behavior theory*. New York: Appleton-Century-Crofts; 1951.
13. Smith AC, Frank LM, Wirth S, Yanike M, Hu D, Kubota Y, *et al.* Dynamic analysis of learning in behavioral experiments. *J Neurosci* 2004; **24**: 447–461.
14. Stefani MR, Moghaddam B. Rule learning and reward contingency are associated with dissociable patterns of dopamine activation in the rat prefrontal cortex, nucleus accumbens, and dorsal striatum. *J Neurosci* 2006; **26**:8810–8818.
15. Sawaguchi T, Goldman-Rakic PS. D1 dopamine receptors in prefrontal cortex: involvement in working memory. *Science* 1991; **251**:947–950.
16. Kiyatkin EA, Rebec GV. Striatal neuronal activity and responsiveness to dopamine and glutamate after selective blockade of D1 and D2 dopamine receptors in freely moving rats. *J Neurosci* 1999; **19**:3954–3609.
17. Baeg EH, Kim YB, Huh K, Mook-Jung I, Kim HT, Jung MW. Dynamics of population code for working memory in the prefrontal cortex. *Neuron* 2003; **40**:177–188.
18. Pan WX, Schmidt R, Wickens JR, Hyland BI. Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. *J Neurosci* 2005; **25**:6235–6242.
19. Sutton RS. Learning to predict by the method of temporal differences. *Mach Learn* 1988; **3**:9–44.
20. Bayer HM, Glimcher PW. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 2005; **47**:129–141.
21. Beninger RJ, Miller R. Dopamine D1-like receptors and reward-related incentive learning. *Neurosci Biobehav Rev* 1998; **22**:335–345.
22. Schmitzer-Torbert N, Redish AD. Neuronal activity in the rodent dorsal striatum in sequential navigation: separation of spatial and reward responses on the multiple T task. *J Neurophysiol* 2004; **91**:2259–2272.
23. Pratt WE, Mizumori SJ. Neurons in rat medial prefrontal cortex show anticipatory rate changes to predictable differential rewards in a spatial memory task. *Behav Brain Res* 2001; **123**:165–183.
24. Dayan P, Abbott LF. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Cambridge: MIT Press; 2001.
25. Izhikevich EM. Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cereb Cortex* 2007; **17**:2443–2452.