Behavioral/Cognitive

Mouse Behavior on the Trial-Unique Nonmatching-to-Location (TUNL) Touchscreen Task Reflects a Mixture of Distinct Working Memory Codes and Response Biases

Daniel Bennett,[1] Jay Nakamura,[2,3] Chitra Vinnakota,[2] Elysia Sokolenko,[4] Jess Nithianantharajah,[5] Maarten van den Buuse,[6] Nigel C. Jones,[7,8,9] Suresh Sundram,[2,10] and Rachel Hill[2]

[1]School of Psychological Sciences, Monash University, Melbourne, Victoria 3180, Australia, [2]Department of Psychiatry, Monash University, Melbourne, Victoria 3180, Australia, [3]Laboratory for Molecular Mechanisms of Brain Development, RIKEN Center for Brain Science, Saitama, Japan, 351-0198, [4]Discipline of Anatomy and Pathology, School of Biomedicine, University of Adelaide, Adelaide, South Australia 5005, Australia, [5]Florey Institute of Neuroscience and Mental Health, Melbourne, Victoria 3052, Australia, [6]School of Psychology and Public Health, La Trobe University, Melbourne, Victoria 3086, Australia, [7]Department of Neuroscience, Central Clinical School, Monash University, Melbourne, Victoria 3004, Australia, [8]Department of Neurology, Alfred Hospital, Commercial Road, Melbourne, Victoria 3004, Australia, [9]Department of Medicine, Royal Melbourne Hospital, University of Melbourne, Melbourne, Victoria 3052, Australia, and [10]Mental Health Program, Monash Health, Clayton, Victoria 3168, Australia

The trial-unique nonmatching to location (TUNL) touchscreen task shows promise as a translational assay of working memory (WM) deficits in rodent models of autism, ADHD, and schizophrenia. However, the low-level neurocognitive processes that drive behavior in the TUNL task have not been fully elucidated. In particular, it is commonly assumed that the TUNL task predominantly measures spatial WM dependent on hippocampal pattern separation, but this proposition has not previously been tested. In this project, we tested this question using computational modeling of behavior from male and female mice performing the TUNL task ($N = 163$ across three datasets; 158,843 trials). Using this approach, we empirically tested whether TUNL behavior solely measured retrospective WM, or whether it was possible to deconstruct behavior into additional neurocognitive subprocesses. Overall, contrary to common assumptions, modeling analyses revealed that behavior on the TUNL task did not primarily reflect retrospective spatial WM. Instead, behavior was best explained as a mixture of response strategies, including both retrospective WM (remembering the spatial location of a previous stimulus) and prospective WM (remembering an anticipated future behavioral response) as well as animal-specific response biases. These results suggest that retrospective spatial WM is just one of a number of cognitive subprocesses that contribute to choice behavior on the TUNL task. We suggest that findings can be understood within a resource-rational framework, and use computational model simulations to propose several task-design principles that we predict will maximize spatial WM and minimize alternative behavioral strategies in the TUNL task.

Key words: computational modeling; hippocampus; prospective working memory; retrospective working memory; TUNL; working memory

Significance Statement

Touchscreen tasks represent a paradigm shift for assessment of cognition in nonhuman animals by automating large-scale behavioral data collection. Their main relevance, however, depends on the assumption of functional equivalence to cognitive domains in humans. The trial-unique, delayed nonmatching to location (TUNL) touchscreen task has revolutionized the study of rodent spatial working memory. However, its assumption of functional equivalence to human spatial working memory is untested. We leveraged previously untapped single-trial TUNL data to uncover a novel set of hierarchically ordered cognitive processes that underlie mouse behavior on this task. The strategies used demonstrate multiple cognitive approaches to a single behavioral outcome and the requirement for more precise task design and sophisticated data analysis in interpreting rodent spatial working memory.

Received Nov. 28, 2022; revised Apr. 28, 2023; accepted June 2, 2023.

Author contributions: D.B. contributed unpublished reagents/analytic tools; D.B. analyzed data; D.B. wrote the first draft of the paper; D.B., J.N., C.V., E.S., J.N., M.v.d.B., N.C.J., S.S., and R.H. edited the paper; D.B., S.S., and R.H. wrote the paper; J.N., C.V., E.S., J.N., M.v.d.B., N.C.J., S.S., and R.H. designed research; J.N., C.V., and E.S. performed research.

This work was supported by the One in Five Foundation; and International Society for Neurochemistry Career Development Grant to R.H.

The authors declare no competing financial interests.

Correspondence should be addressed to Daniel Bennett at daniel.bennett@monash.edu or Rachel Hill at rachel.hill@monash.edu.

https://doi.org/10.1523/JNEUROSCI.2101-22.2023

Introduction

Working memory (WM), the active maintenance and manipulation of a small amount of information within a transient, limited-capacity memory store, is fundamental to many tasks that humans face in their day-to-day lives (Baddeley, 2010; Cowan, 2014; Oberauer and Lin, 2017; Gold and Luck, 2022). WM is impaired in many psychiatric disorders (Steele et al., 2007; Kofler et al., 2018; Gold et al., 2019), and WM deficits are predictive of poor clinical and functional outcomes in both schizophrenia and autism (Troyb et al., 2014; Fu et al., 2017). Despite this, there are no current treatments that address WM impairments in psychiatric disorders. In overcoming this absence, animal behavioral testing paradigms are key to trialing the effects of novel therapeutic compounds on cognition (Castner et al., 2004; Dudchenko et al., 2013). However, the translation of novel therapeutics to clinical application is dependent on the validity of the animal behavioral paradigms used to assay cognitive deficits in preclinical evaluations of potential therapeutics (Pound and Ritskes-Hoitinga, 2018).

Behaviors indicative of WM have been observed across insects, fish, birds, rodents, and nonhuman primates (Roberts, 1972; Miller et al., 1996; Aultman and Moghaddam, 2001; Giurfa et al., 2001; Bloch et al., 2019). This similarity notwithstanding, however, there are pronounced between-species differences in WM, particularly in memory capacity (i.e., in the number of representations that can be held in WM) and in the rate of forgetting during a retention interval (Dudchenko, 2004; Carruthers, 2013; Lind et al., 2015; Roberts and Santi, 2017). From a translational perspective, this raises the concern that neurocognitive processes of WM in a given species may be too dissimilar to human WM to use the species as a translational model. Recent initiatives have sought to address this concern by developing animal WM tasks that are maximally similar to standard human WM tasks (Barch et al., 2009, 2012; Dudchenko et al., 2013), thereby allowing researchers to assay WM processes that are similar across species despite quantitative between-species differences in the capacity and forgetting rate of WM. One particularly influential task is the trial-unique nonmatching-to-location (TUNL) task, a touchscreen-based nonmatch-to-sample task that allows for high-throughput measurement of spatial WM in rodents (Talpos et al., 2010; Bussey et al., 2012; Oomen et al., 2013; Kim et al., 2015).

An important unresolved question concerns the representational code with which information is held in rodent WM during the TUNL task. Broadly, we can distinguish a retrospective code, in which WM maintains a representation of a previous sample stimulus, from a prospective code, in which WM maintains a representation of a planned behavioral response (see Roitblat, 1982; Cook et al., 1985). Previous work using the TUNL task (e.g., Talpos et al., 2010; Kim et al., 2015; Zeleznikow-Johnston et al., 2017) has tended to interpret behavioral results as reflecting spatial WM for the location of a previous sample stimulus, that is, a retrospective WM code for spatial location, rather than a planned future behavioral response. It is indeed the case that many standard human WM tasks assess retrospective WM (e.g., Hopkins et al., 1995; Della Sala et al., 1999; Nunn et al., 1999). However, nonhuman animals have been shown to use a mixture of prospective and retrospective memory codes across tasks, including 12-arm radial mazes, plus mazes, and delayed match-to-sample tasks (Cook et al., 1985; Kametani and Kesner, 1989; Kesner, 1989; Ferbinteanu and Shapiro, 2003), and humans may also use a mixture of retrospective and prospective WM under some circumstances (Zimmer, 2008).

Understanding WM coding within the TUNL task is also crucial for the neural interpretation of TUNL data. Under the assumption that it reflects retrospective WM for spatial location, TUNL behavior on small-separation trials is frequently taken as an index of pattern separation algorithms in the hippocampus (Talpos et al., 2010; McAllister et al., 2013; Kumar et al., 2015; Kenton et al., 2018). However, because of the strong links between spatial stimulus encoding and the hippocampus (e.g., Spellman et al., 2015), this neural-algorithmic interpretation of TUNL data depends on the assumption that animals are solely using retrospective (i.e., spatial) WM. The reason for this is that, by definition, pattern separation can only be an explanation of behavior in circumstances where the two stimuli being pattern-separated by the hippocampus are similar to one another in WM. For small-separation trials in the TUNL, the two stimuli are only similar to one another in memory if we assume that it is their spatial location that is being encoded in WM (in other words, a retrospective WM code).

The nature of WM coding in the TUNL task is a crucial question both for the task's translational validity and for the neural interpretation of TUNL behavior. In the present study, we sought to rigorously test the relative contributions of retrospective and prospective WM to TUNL behavior (as well as several response biases unrelated to memory), using hierarchical Bayesian modeling of single-trial behavioral data from three distinct datasets of mice completing the TUNL task. Based on previous studies detailing usage of both retrospective and prospective WM in rodents (e.g., Kametani and Kesner, 1989; Kesner, 1989), we hypothesized that mouse behavior on the TUNL task might be driven by both retrospective and prospective WM.

Materials and Methods

Overview of TUNL task

The TUNL task is a touchscreen spatial nonmatch-to-sample task that was designed as an assay of rodent WM (Talpos et al., 2010; Bussey et al., 2012; Oomen et al., 2013). In each trial, mice within a touchscreen operant chamber must first touch an illuminated "sample" location (one of five horizontally spaced locations on the touchscreen, signaled by increased luminance; see Fig. 1*A*). After a delay (the "retention interval") in which no locations are illuminated, animals complete a choice phase in which two locations are illuminated: one that matches the sample location and one that is nonmatching. Animals receive a reward if they touch the nonmatching location.

Trial difficulty in the TUNL task is typically manipulated (see, e.g., Talpos et al., 2010; Bussey et al., 2012; Oomen et al., 2013; Kim et al., 2015) either by increasing the duration of the retention interval (thereby increasing the duration for which the sample location must be held in WM) or by reducing the spatial separation between the sample location and the nonmatching location (thereby increasing the similarity between the two possible choice options). Different separation conditions in the TUNL task are typically described based on the number of unlit locations that separate the two response options in the choice phase (e.g., the example trial in Fig. 1*A* is referred to as a Separation-3 or S3 trial because three unlit squares separate the two response options in the choice phase).

Each separation condition in the TUNL task can be constructed using multiple different configurations of sample location and nonmatch location: a Separation-3 (or S3) trial, for instance, can be constructed either with the sample at the far left and the (correct) nonmatch option at the far right, or vice versa. This is a strength of the task because, unlike earlier nonmatch-to-sample tasks involving only two response levers (e.g., Chudasama and Muir, 1997), there are 20 different unique configurations of sample and nonmatching location that can be tested across the four different separation conditions (see Fig. 1*B*). A relative increase in trial uniqueness is thought to reduce the likelihood of mediating behavioral responses (e.g., repeatedly pressing a response lever during the

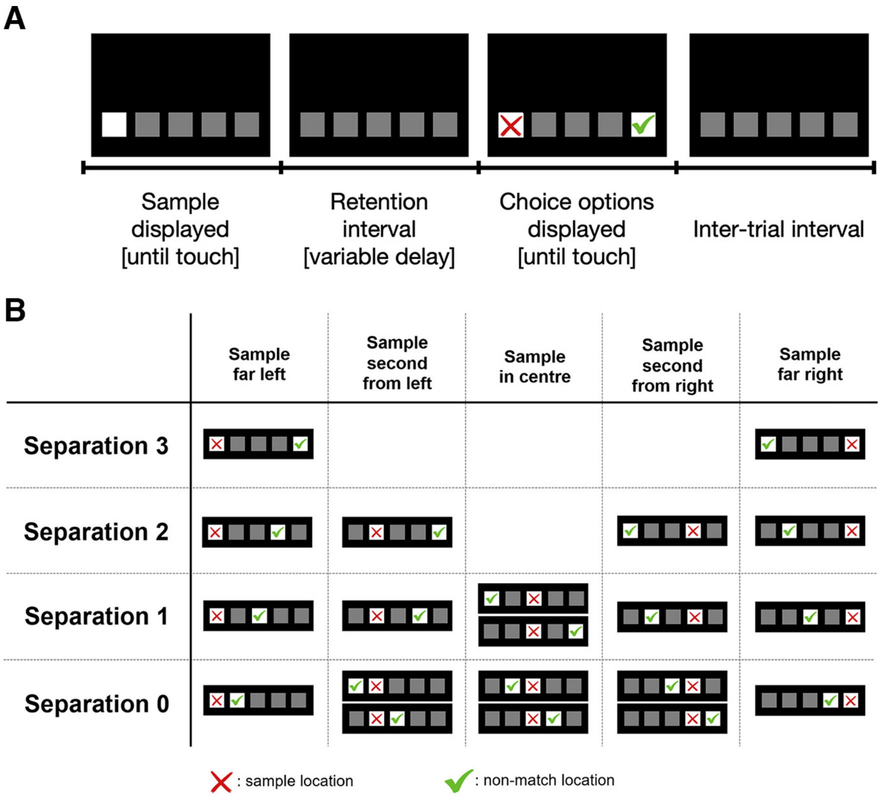

Figure 1. Trial schematic for the mouse TUNL task. ***A***, Sequence of events in each trial. The animal initiates the delay period by touching the sample location (white square). After the retention interval, the animal is shown two lit squares; to receive a reward, it must touch the square that is a nonmatch for the trial's sample location (green tick represents correct location; red cross represents incorrect location; ticks and crosses are for illustration purposes only and are not visible to mice). ***B***, Overview of each of the 20 possible configurations of sample and nonmatch location. Separation conditions are defined in terms of the number of unlit squares that separate the sample location and the nonmatching location (e.g., in a Separation 3 trial, the sample location and the nonmatch location are separated by three unlit squares). Within each separation type, there are multiple different configurations of sample locations (red crosses) and nonmatch locations (green ticks). The 20 different possible configurations of sample location and nonmatch location that can be presented comprise two configurations constituting a Separation 3 trial, four for Separation 2 trials, six for Separation 1 trials, and eight for Separation 0 trials. Simulated proportions of correct responses for each separation condition under retrospective and prospective WM can be found in Extended Data Figure 1-1.

retention interval) (Talpos et al., 2010). This feature of the TUNL task therefore also increases the similarity between the TUNL task and comparable visuo-spatial WM tasks in humans that also rely on remembering the spatial location or configuration of stimuli (e.g., Park and Holzman, 1992; Della Sala et al., 1999; Duff and Hampson, 2001; for review, see Logie, 2014).

In the TUNL task, an animal's ability to respond at the nonmatching location in the choice phase is typically taken as an indication that the animal has retained the spatial location of the sample stimulus in its WM throughout the retention interval. For this reason, the TUNL task is considered a task of (retrospectively coded) spatial WM. However, it is also possible for mice to attain above-chance performance on the TUNL task using prospective WM alone (although the maximum performance of prospective WM is still inferior to performance levels achievable using retrospective WM; see Extended Data Fig. 1-1). This can be done by using the sample stimulus to encode, not a spatial location, but a behavioral intention to respond either at the leftmost or the rightmost of the response options in the choice phase. For trials with a sample on the far left, for example, accuracy would be 100% for a mouse that simply choose the rightmost of the two choice options (see Fig. 1*B*, leftmost column). Extended Data Figure 1-1 provides more information of the respective utilities of retrospective and prospective WM on the TUNL task; it is important to note that, despite the sophistication of the TUNL task design, it is nevertheless possible for animals to achieve above-chance performance in this task using only prospective WM.

Overview of datasets

The present study reports data from a total sample of 158,843 choice trials from $N = 163$ mice completing the TUNL task in three distinct datasets (two of which have been previously published; see Table 1). We focused our analyses on the post-training "probe" test phase of the task, after animals had successfully learned the task to asymptote. It is behavior on these probe trials that is typically taken as the measure of spatial WM performance in studies using the TUNL task (e.g., Kim et al., 2015; Nilsson et al., 2016; Zeleznikow-Johnston et al., 2017).

Although the three datasets had equivalent training protocols, each dataset used a slightly different set of trial types in the probe phase of the task (as is common practice in studies using the TUNL task). For example, in the probe phase of Datasets 2 and 3 (but not the probe phase of Dataset 1), incorrect responses were followed by "correction trials," in which the same two choice locations remained illuminated until animals correctly selected the nonmatching location. As per standard training protocols, correction trials were provided during the training phase for all three datasets. In choosing datasets for analysis, we deliberately selected datasets with a heterogeneous set of trial types because our overall goal was to ensure that our model selection procedures were broadly generalizable. By fitting models to datasets that differed in their trial types and probe-phase parameters, our goal is to avoid overfitting our models to one specific set of trial types or testing parameters. If we observe that the same computational model provides the best fit to data across datasets despite the large differences between datasets, we can have increased confidence that the model applies to TUNL behavior broadly, rather than simply TUNL behavior for specific testing parameters. Similarly, by testing across animal cohorts that were genetically diverse and that received different experimental manipulations (see below), we sought to ensure that results were generalizable across different animal and experiment types.

Table 1. Overview of datasets[a]

Dataset #	Original publication	*N* mice (female/male)	*N* probe-phase trials	Separation conditions	Delays (s)	Correction trials at test?
1	Nakamura et al. (2021)	83 (46/37)	113 599	S0, S1, S2, S3	0, 3, 6, 9, 12, 15, 18, 21, 24	No
2	Vinnakota et al. (in preparation)	44 (19/25)	24 947	S1, S2, S3	1, 2, 3	Yes
3	Sokolenko et al. (2020)	36 (0/36)	20 297	S1[b]	1, 2, 3, 6	Yes

[a] For details of the training protocol for all datasets before the probe phase, see Extended Data Table 1-1.

[b] Dataset 3 only tested S1 trials for which either the sample location or the nonmatch location was at the center response location (Fig. 1B: Separation-1 row, columns 1, 3, and 5).

All animal experiments were conducted in accordance with the guidelines in the Australian Code of Practice for the Care and Use of Animals for Scientific Purposes (National Health and Medical Research Council of Australia, Ed 8, 2013) and the ARRIVE guidelines. Mice from Dataset 1 were C57Bl/6J mice obtained from the breeding colony at the Monash Animal Research Platform, Monash Medical Center (Clayton, Victoria, Australia). All subsequent husbandry, housing, and behavioral testing were also undertaken at Monash University. Mice were exposed prenatally to maternal immune activation manipulation, such that timed mated dams were injected either with polyinosinic-polycytidylic (5 mg/kg) solution or vehicle saline solution via the intraperitoneal route (10 ml/kg) on 3 consecutive days of GD 9,10, and 11 or GD 13, 14, or 15. Offspring were weaned at postnatal day 21. Mice were housed in a reverse 12:12 h light cycle in individually ventilated cages (GM500, Tecniplast) with *ad libitum* access to food and water until the initiation of food restriction. Cages were monitored daily and changed fortnightly. This experiment was approved under Monash University Animal Ethics Committee MMCB/2017/10. Further details regarding this dataset are available in Nakamura et al. (2021).

For Dataset 2, mice were transported from Tokyo Metropolitan Institute of Medical Science to a breeding colony established and maintained at the Monash Animal Research Platform, Monash Medical Center. GluN2D heterozygous male and female C57BL/6J mice were bred to obtain WT, heterozygous, and homozygous GluN2D-KO littermates. All mice were housed in groups of 2-5 in individually ventilated cages (Tecniplast) with *ad libitum* access to food and water until the initiation of food restriction. At 6-7 weeks of age, mice were transferred from the breeding facility to the behavioral holding room with a reversed 12 h dark-light cycle (lights off at 8:30 A.M.) and kept there until the end of the experiment. Cages were monitored daily and changed fortnightly. Male and female WT and GluN2D-KO mice were used in behavioral testing. This experiment was approved under Monash University Animal Ethics Committee (#E/1837/2018/M).

C57Bl/6 in Dataset 3 PV-Cre (JAX: 008069) and CaMKIIα-Cre (JAX: 005359) mice were crossed with GluN1-floxed mice (JAX: 005246) to obtain either PV$^+$ interneuron-specific GluN1 KO mice (PV-Cre/wt; GluN1fl/fl), or forebrain pyramidal cell-specific GluN1 KO mice (CaMKIIα-Cre/wt; GluN1fl/fl), or WT littermates controls (w/w; GluN1fl/fl or w/w; GluN1fl/wt). Mice were originally obtained from The Jackson Laboratory, and group-housed in open top cages in the Kenneth Myer Building, Parkville, Victoria, Australia, on a reverse-lighting schedule (lights off from 0700 to 1900). Mice were provided *ad libitum* access to standard chow until 10 weeks of age, when food restriction was initiated. This experiment was approved under the Florey Neuroscience Institute Animal Ethics Committee (#16-028). Further details regarding this dataset are available in Sokolenko et al. (2020).

Before touchscreen testing, mice were gradually food restricted over a period of ∼3-5 d until reaching 85%-90% of their free-feeding weight, which was maintained until the end of testing. Strawberry milk (Nippy's) liquid food reward was introduced 2 d before testing to familiarize the mice with the reward. Mice were handled daily for 1 week before initiating touchscreen training to habituate to the handler. The TUNL task was run in isolated touchscreen operant chambers for mice (Campden Instruments) through ABET II software. Chambers were dedicated to either male or female mice only. The apparatus was cleaned with ethanol 80% v/v after each session. The training schedule used to train mice in use of the apparatus and in performance of the TUNL tasks is provided in Extended Data Table 1-1. In the present study, we solely analyzed data drawn from blocks of the task after animals had learned the task to criterion (i.e., from the "probe" phase of the task). Within the probe phase of the task, delay manipulations were fixed for all datasets (i.e., each probe testing session assessed performance on a single delay length). In Dataset 1, testing sessions each assessed a mixture of different separation conditions, whereas in Datasets 2 and 3, each testing session assessed a single separation condition.

All mice were trained to criterion on the task using a standard habituation and conditioning protocol (Kim et al., 2015), with minor variation between datasets in the criterion asymptotic performance level required to proceed to the testing phase (Dataset 1: accuracy ≥70%; Dataset 2: accuracy ≥80%; Dataset 3: accuracy ≥75%). All other training parameters were identical to those described by Kim et al. (2015), with the exception that mice in Datasets 2 and 3 were not trained to criterion on S0 trials.

In all datasets, we report analyses from the entire sample of mice without reference to distinctions between experimental groups in the original publications (e.g., maternal immune activation for a subset of mice in Dataset 1). This was because our goal was to assess patterns of behavior that were observed consistently across multiple datasets of TUNL data from different samples of mice, without reference to distinctions between groups of mice in any specific experiment. Nevertheless, to ensure that our results were not driven by impaired performance induced by atypical mouse genotypes or experimental manipulations, we repeated all analyses among control mice only. Of particular note, we investigated performance among the WT control mice in Dataset 2, which were genetically typical and did not undergo any form of experimental manipulation which might have been expected to impair performance The results of these analyses indicated that similar overall behavioral patterns were observed even when we restricted analyses to control mice only (see Extended Data Table 2-5; Extended Data Fig. 2-1). To maximize statistical power and emphasize the generalizability of our results, we therefore report analyses from the entire sample of mice in each dataset in this manuscript.

Experimental design and statistical analysis

For statistical analysis of TUNL data, we adopted a two-stage approach. First, we used a series of model-agnostic analyses (detailed immediately below) to quantify relevant aspects of animals' choice behavior using standard inferential statistics. We then conducted a series of hierarchical Bayesian computational modeling analyses to address more nuanced questions about the relative strength of the effects of different WM codes and response biases on task behavior.

Logistic regression analyses. The effects of separation, delay, and trial type (i.e., the 20 unique configurations as per Fig. 1*B*) on behavior on the TUNL task were assessed using mixed-effects logistic regression analyses, with choice accuracy as the dependent variable (choice of nonmatch location coded as 1, choice of sample location coded as 0). Analyses were conducted using the *lme4* package for R (Bates et al., 2015). All regression models included per-animal random intercepts, as well as per-animal slopes for all main effects and interactions that were entirely within-animal (Barr et al., 2013). Nonconverging models were simplified by removing higher-order random slopes until convergence was achieved. Coefficient *p* values were calculated using the Wald *t*-to-*z* test (Meteyard and Davies, 2020).

Calculation of side bias metric. We computed a measure of the strength of each animal's preference for responding in a leftward/rightward direction, taking into account the different choice options available within the different datasets. To do this, we first calculated the choice disparity for each of the five possible response locations, which is a

Table 2. Overview of computational model fits to each dataset[a]

Model number	N model parameters per mouse	Forgetting function	Retrospective WM	Prospective WM	Side biases	Distal-response biases	Dataset 1		Dataset 2		Dataset 3	
							WAIC	ΔWAIC (SE)	WAIC	ΔWAIC (SE)	WAIC	ΔWAIC (SE)
M1	3[b]	Power-law	✓	—	—	—	71 495.2	15 678.6 (362.0)	7940.5	584.8 (74.3)	3807.6	441.9 (76.1)
M2	2	Power-law	—	✓	—	—	71 471.2	15 654.7 (361.1)	8491.5	1135.8 (103.5)	6753.4	3387.6 (359.0)
M3	1	NA	—	—	✓	—	66 937.5	11 120.9 (370.9)	14 189.0	6833.3 (300.2)	10 405.6	7039.8 (576.7)
M4	1	NA	—	—	—	✓	76 250.8	20 434.3 (422.7)	14 321.4	6965.7 (296.0)	10 502.6	7136.8 (571.0)
M5	4	Power-law	—	✓	✓	✓	56 025.8	211.3 (65.5)	7378.2	22.5 (15.7)	3812.1	446.3 (80.9)
M6	5[b]	Power-law	✓	—	✓	✓	57 088.1	1271.6 (117.6)	7511.2	155.5 (36.7)	3382.2	16.5 (18.5)
M7	5[b]	Power-law	✓	✓	—	✓	68 261.1	12 444.5 (335.2)	7771.4	415.7 (61.4)	3707.8	342.0 (69.7)
M8	5[b]	Power-law	✓	✓	✓	—	58 258.4	2441.9 (148.0)	7511.9	156.2 (36.7)	3401.1	35.3 (17.4)
M9	6[b]	Power-law	✓	✓	✓	✓	55 816.5	—	7355.0	—	3365.8	—

[a] WAIC values are presented on a deviance scale (lower values indicate better model fit). Equivalent model fits for exponential and sigmoidal forgetting functions can be found in Extended Data Tables 2-1 and 2-2, respectively. Model parameter estimates for the best-fitting model M9 can be found in Extended Data Table 2-3. Results of a further comparison of variants of M9 with two forgetting rates can be found in Extended Data Table 2-4. Results of model comparison for data from control animals only can be found in Extended Data Table 2-5.

[b] There is one fewer parameter per mouse in Dataset 3 because a smaller number of separations in this dataset rendered the α parameter nonidentifiable.

normalized measure of how often each animal responded at each response location (compare Broschard et al., 2021) as follows:

$$\text{Disparity}(i) = \frac{N_{chosen}(i) - N_{correct}(i)}{N_{total}} \tag{1}$$

Here, $i \in [1, 2, 3, 4, 5]$ denotes the response location (1 = far left, 3 = center, 5 = far right), $N_{chosen}(i)$ denotes the number of times that the animal responded at location *i*, and $N_{correct}(i)$ denotes the number of times that location *i* was the correct response location. The denominator N_{total} denotes the total number of trials completed by the animal, allowing us to compare the choice disparity index across animals who completed different total numbers of trials. This metric can be interpreted as an index of whether an animal chose a particular response location more often than expected given the trial types it completed (positive disparity), or less often than expected (negative disparity).

For each animal, we then computed the side bias index simply by calculating the difference between the choice disparity for rightward response options and the choice disparity for leftward response options. This provides a normalized index of whether an animal preferred rightward response options (positive side bias metric) or leftward response options (negative metric) as follows:

$$\begin{aligned}\text{Side bias } &= \big(\text{Disparity}(4) + \text{Disparity}(5)\big) \\ &\quad - \big(\text{Disparity}(1) + \text{Disparity}(2)\big)\end{aligned} \tag{2}$$

Within Figure 3, we computed the animal-by-animal significance of the side bias metric using a nonparametric empirical permutation test. For each animal, we tested significance by randomly shuffling the actual N_{chosen} vector 1000 times to estimate an empirical null distribution of the side bias metric. An animal's side bias was taken to be significantly different from chance if the actual metric as computed by Equation 2 fell outside the 95% CI of this metric as estimated by the empirical permutation test.

Calculation of distal-response bias metric. Similar to the side bias metric above, we also computed a measure of the strength of each animal's preference for responding at locations closer to the edges of the testing arena ("distal" responses) versus locations in the center of the testing arena ("central" responses options). The computation of the distal-response bias metric was also based on differences in choice disparities as defined in Equation 1. In this case, however, we were solely concerned with whether each animal responded at the central response location more or less often than would be expected by chance as follows:

$$\text{Centre bias} = \text{Disparity} \tag{3}$$

This computes an index of whether animals preferred the central response option or more distal response options (positive/negative values, respectively). As with the side bias index, significance was estimated on a per-animal basis using a nonparametric empirical permutation test.

Computational models

The model-agnostic statistical analyses described above are appropriate for answering high-level questions about the patterns evident in the behavioral data. For a more nuanced consideration of the relative contributions of different WM coding schemes and response biases to task behavior, we next turned to computational modeling within a hierarchical Bayesian framework. Our overarching goal in this analysis was to determine the extent to which behavior was determined by four separable response factors:

- *Retrospective WM:* memory for the spatial location of a previous sample stimulus
- *Prospective WM:* using the sample stimulus to encode an intended response direction in the choice phase
- *Side biases*: animal-specific preferences for responding in either a leftward or a rightward direction, independent of the sample stimulus (i.e., no WM component)
- *Distal-response biases*: animal-specific preferences for responding at locations either closer to the walls of the chamber or closer to the center of the chamber, independent of the sample stimulus (i.e., no WM component)

All models assumed that choices on a given trial were driven by the competition between two response strengths: one for the matching (i.e., incorrect) location and one for the nonmatching (i.e., correct) location (compare Nosofsky, 1986; Kruschke, 1992). We compared a set of different computational models that differed in their assumptions about how the response strengths for each response location was determined. The specific set of models that we compared comprised all one-, three-, and four-way combinations of the four response factors (see Table 2). As such, formal comparison of computational models allowed us to determine which response factors were important in driving behavior, and estimation of model parameters allowed us to quantify the strength of the effects of each response factor.

Modeling framework

In all models, the response strengths for the matching and nonmatching response locations were denoted R_m and R_{nm}, respectively. The probability of making a (correct) nonmatch response was assumed to be a logistic function of the difference between these competing response strengths (also known as a softmax function or Luce choice rule) on a given trial *t* as follows:

$$\Pr(nm) = \frac{1}{1 + e^{R_m(t) - R_{nm}(t)}} \tag{4}$$

Different models made contrasting assumptions about how these response strengths were calculated. Specifically, different models comprised different additive combinations of four different response factors: retrospective WM, prospective WM, side biases, and distal-responses biases. For instance, in the most complex model (model M9) all four response factors influenced choice, and response strengths were calculated as per Equation 5 as follows:

$$R_{loc} = w_{retro} Retro(loc, d) + w_{prosp} Prosp(loc, d) + w_{side} Side(loc) + w_{distal} Distal(loc) \quad (5)$$

Where *loc* stands for either the sample location *m* or the nonmatching location *nm* ($m, nm \in [1, 2, 3, 4, 5]$), and *d* denotes the duration of the retention interval. Each of the *w* parameters in Equation 5 is an animal- and component-specific weighting parameter that controls the strength of each of the response factors (as specified by the functions *Retro*, *Prosp*, *Side*, and *Distal*; see below). By contrast, the simpler models M1-M8 consisted of more restricted combinations of the different combinations, as specified in Table 2. For instance, response strengths in the retrospective-memory-only M1 were computed as per Equations 6 and 7 as follows:

$$R_{loc} = w_{retro} Retro(loc, d) \quad (6)$$

In all models, the key distinction between the WM factors (*Retro* and *Prosp* functions) and the response-bias factors (*Side* and *Distal*) was that the response strength produced by WM factors was assumed to decrease over time as information was forgotten from WM during the retention interval of each trial (i.e., these functions depend both on the locations *m* and *nm* and the delay *d*), whereas the response strength produced by response biases was assumed to remain constant over time within each trial (and therefore depend only on *m* and *nm*). In all models reported in the main text, we modeled the rate of this information loss using a power-law forgetting function (Wickelgren, 1974; Wixted and Carpenter, 2007; Donkin and Nosofsky, 2012) as follows:

$$Retro(loc, d) = Retro^0(loc)(1 + d)^{-\beta} \quad (7a)$$

$$Prosp(loc, d) = Prosp^0(loc)(1 + d)^{-\beta} \quad (7b)$$

Here, the remaining response strength produced by WM after *t* seconds depends both on the initial response strength functions $Retro^0$ and $Prosp^0$ (i.e., the response strength immediately after the presentation of the sample location as per retrospective and prospective WM, respectively; see below) and the retention duration *d*. The free parameter β controls the rate at which information is forgotten from WM.[1]

We also tested several alternative forgetting functions (exponential and logistic/sigmoidal functions; Eqs. 8 and 9, respectively). However, we found the power-law function provided the best overall fit to data, and the rank order of different models' goodness of fit was largely unchanged across the different forgetting functions (see Extended Data Tables 2-1 and 2-2). We therefore solely report conclusions from power-law computational models in the main text.

$$Retro(loc, d) = Retro^0(loc) e^{-\beta d} \quad (8a)$$

$$Prosp(loc, d) = Prosp^0(direction) e^{-\beta d} \quad (8b)$$

$$Retro(loc, d) = \frac{Retro^0(loc)}{1 + e^{b(d - mid)}} \quad (9a)$$

$$Prosp(loc, d) = \frac{Prosp^0(loc)}{1 + e^{b(d - mid)}} \quad (9b)$$

Retrospective WM. The retrospective WM equation instantiated the hypothesis that animals made choices in the TUNL task by holding the location of the original sample stimulus in WM during the retention interval. At the choice phase, each possible response location was then assumed to accrue response strength in proportion to its spatial distance to the sample location (i.e., greater response strength for locations at a greater distance from the sample location). We modeled this accrual of response strength using an exponential similarity kernel (Shepard, 1957, 1987) as follows:

$$Retro^0(loc) = e^{\alpha \times |loc - sample|} \quad (10)$$

Where α is a free parameter corresponding to the kernel width of the dissimilarity kernel. This parameter can be interpreted as the degree of cognitive dissimilarity between different response locations (i.e., cognitive capacity for pattern separation). When α is large, different response locations are sharply distinguished from one another; as α approaches 0, the animal treats all possible response locations as equivalent. To avoid parameter nonidentifiability, the α parameter was constrained to be positive in all datasets. α was fixed to a value of 1 in Dataset 3 because the paucity of different separation conditions in this dataset (see Table 1) meant that the α parameter was not uniquely identified.

Prospective WM. The prospective WM equation instantiated the hypothesis that animals made choices in the TUNL task by holding an intended response direction in WM during the retention interval. That is, on seeing the sample location, animals were assumed to form a prospective behavioral intention to select either the leftmost or the rightmost response location at the choice phase (according to whichever direction was most associated with reward in training for each sample location; see Table 1). Specifically, this model component assumed that the response strength for the different locations at the choice phase was proportional to the probability that that response direction would be correct given the sample location as follows:

$$Prosp^0(direction) = \Pr(direction = correct | sample) \quad (11)$$

These probabilities were assumed to be learned by animals throughout the conditioning phases of the task.

Distal-response bias. This response factor assumed that, independent of all other task factors, individual animals idiosyncratically preferred to respond either closer to the walls of the testing chamber (both left and right walls), or to respond closer to the center of the testing chamber as follows:

$$Distal(loc) = \begin{cases} 1, loc = 3 \\ 0, otherwise \end{cases} \quad (12)$$

Where $loc = 3$ denotes that the response location was in the center of the five possible locations. Accordingly, a positive value of the weighting parameter w_{distal} denotes a preference for central stimuli, a negative value denotes a preference for distal stimuli, and a value of zero denotes no overall preference for central versus distal response options.

Side bias. This response factor assumed that, independent of all other task factors, individual animals idiosyncratically preferred to choose either the leftward or the rightward of the two response options on a given trial as follows:

$$Side(loc) = \begin{cases} 1, loc > alt \\ 0, loc < alt \end{cases} \quad (13)$$

Where $loc > alt$ indicates that the response location was to the right of the alternative location, and $loc < alt$ indicates that it was to the right. Accordingly, a positive value of the weighting parameter w_{side} indicates a preference for rightward stimuli, a negative value indicates a preference for leftward stimuli, and a value of zero indicates no bias in either direction.

Model fitting and comparison. All models were fit within a hierarchical Bayesian framework using the probabilistic programming language

[1] In the additional models reported in Extended Data Table 2-4, different β parameters were fit for Equation 7a and Equation 7b for each animal. In all other models, a single β per animal was fit for both Equations 7a and 7b.

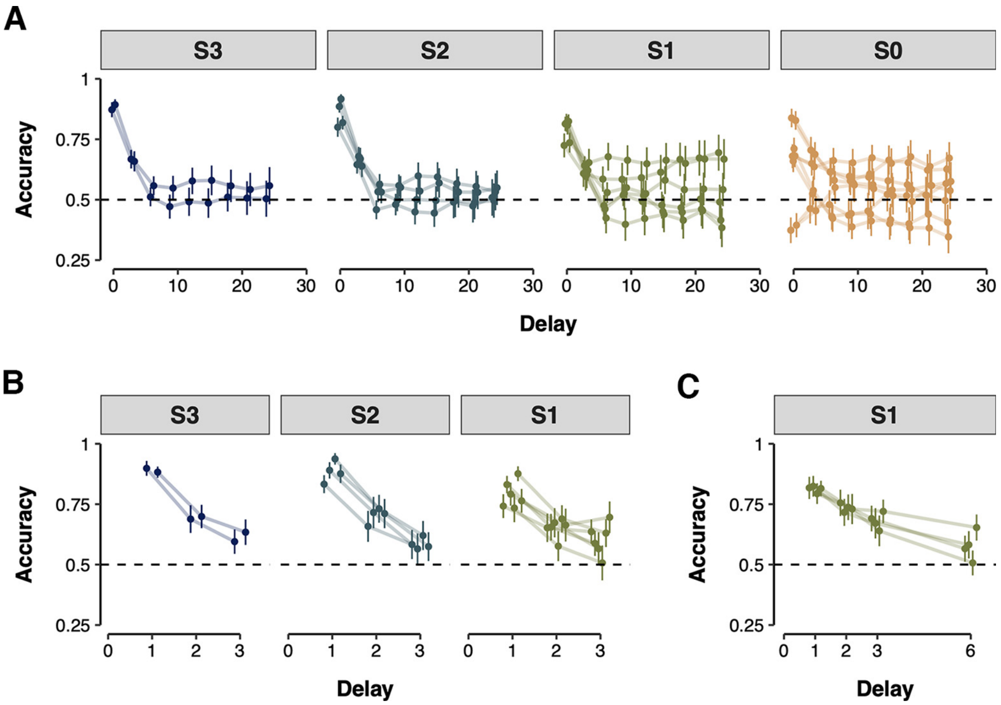

Figure 2. Mouse behavioral data as a function of separation and delay across three independent datasets: ***A***, Nakamura et al. (2021). ***B***, Vinnakota et al. (in preparation). ***C***, Sokolenko et al. (2020). Each subplot represents mean proportion of correct responses (*y* axis) across animals as a function of delay between memory sample and response screen (*x* axis) and separation condition (plot facets). Each individual point within a delay/separation condition presents data from a unique configuration of sample location and nonmatch location (see Fig. 1*B*). Error bars indicate the 95% CI of the mean. Equivalent plots among control mice only can be found in Extended Data Figure 2-1.

Stan (Carpenter et al., 2017) and the *cmdstanr* package for R. For each model, we took 3000 samples from the joint posterior per chain across a total of four chains. The first 1750 samples per chain were discarded to prevent dependence on initial values, resulting in a total of 5000 post-warmup posterior samples being retained for analysis across the four chains. Models were compared using the Watanabe-Akaike Information Criterion (WAIC) (Watanabe, 2013) as estimated within the *loo* package in R (Vehtari et al., 2017). This criterion is an approximation of the leave-one-out posterior predictive density of the data given the model, and assesses goodness of fit while penalizing models with excessive complexity. Statistical ties between different models (i.e., differences in WAIC of <1 SE of the difference of WAIC) were broken according to statistical parsimony, defined as number of parameters per animal (Table 2). All chains converged in all models as indicated by an $\hat{R}$ value <1.1, and there were no divergent transitions in any model. All models used partial pooling, with animal-level parameters assumed to be drawn from a group-level distribution with a mean and SD estimated from the data. Point estimates of animal-level parameters were calculated as the median of the posterior distribution per parameter per animal.

Variance partitioning analysis. Once we had identified a best-fitting model according to the procedure above, we next computed the amount of variance explained by each response factor in that model. In this analysis, we calculated the change in the overall group-level model R^2 when different response factors were removed one at a time from the best-fitting model. R^2 values in this analysis were computed for each trial type and delay duration at the group-mean level.

Model simulations. After selecting the best-fitting model overall, we conducted two sets of simulation analyses to dissect the behavioral signatures of retrospective versus prospective WM in the TUNL task.

First, we conducted several simple simulations to calculate the expected proportion correct for different configurations of sample location and nonmatch location under retrospective WM versus prospective WM (see Fig. 1*B*; Extended Data Fig. 1-1). In these simulations, we assumed that each of the 20 trial types in Figure 1*B* was presented equally often, and that animals deterministically selected the response locations indicated by either a retrospective or prospective WM code (or selected at random between the different indicated locations in cases where more than one response location was equally indicated). For retrospective WM, simulations assumed that animals were perfectly able to select the nonmatching location. For prospective WM, simulations assumed that animals selected a response direction (leftward vs rightward) on the basis of the sample location alone, and then deterministically responded at the location that was further in the selected response direction in the choice phase. The response direction was assumed to be learned during the conditioning phase of the task by marginalizing across the different subsequent nonmatch locations that followed each possible sample location. For instance, when the sample location was second from the left (see Fig. 1*B*, second column), then 75% of the time the correct response location would be the rightward of the two options in the choice phase. The simulations reported in Extended Data Figure 1-1 therefore assume that mice deterministically responded at the rightward response location after a sample location that was second from the left (and were therefore correct on 75% of trials for this sample location).

The second set of simulations relaxed the unrealistic assumption that animals were able to perfectly execute either retrospective or prospective WM. Here, we estimated the performance of both retrospective and prospective WM under the assumption that the response strength according to WM (of either kind) became weaker over time, consistent with the behavioral data presented in Figure 2. For each dataset, we specifically estimated the performance of each WM code using point estimates of the forgetting rate parameter β for the best-fitting model for that dataset. Since these simulations were probabilistic rather than deterministic, we repeated simulations 10,000 times and computed the proportion of datasets in which an animal using retrospective WM alone would obtain more reward than an animal using prospective WM alone. For each dataset, we simulated performance according to the actual configuration of trial types that were presented in that dataset. To inform the design of future studies, we also simulated performance under other trial configurations that experimenters might wish to test (Extended Data Fig. 6-1).

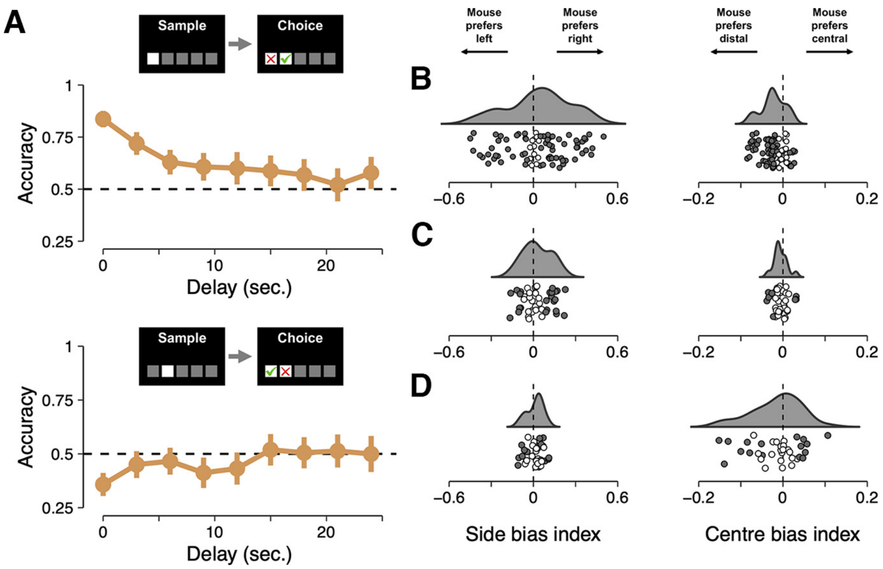

Figure 3. ***A***, Demonstration of direction-dependent effects in two putatively similar S0 choice configurations across multiple delays from Dataset 1. Top, Sample location (red cross) at far left and nonmatch location (green tick) second from the left; mice display above-chance performance at delay 0 that deteriorates with increasing delay duration. Bottom, Sample location second from the left and nonmatch location at the far left; mice show below-chance performance at delay 0, and improving performance as delays increase. Error bars indicate the 95% CI of the mean. Equivalent results for sample locations at far right and second from right can be found in Extended Data Figure 3-1. ***B–D***, Side biases (left column) and center-biases (right column) for Dataset 1 (***B***), Dataset 2 (***C***), and Dataset 3 (***D***). A positive side bias index indicates that a mouse preferred to respond at the right-hand side of the chamber, and a negative side bias index indicates that a mouse preferred to respond at the left-hand side of the chamber. A positive center bias indicates that a mouse preferred to respond at the center of the chamber, and a negative center bias indicates that a mouse preferred to respond at the edges of the chamber. Distributions represent estimates of the group-level distributions in each dataset. Individual points represent statistics for individual animals. Filled data points represent animals with biases that were significantly different from chance. Unfilled data points represent animals with biases not significantly different from chance.

Results

Different configurations at the same spatial separation are not interchangeable

In previous studies, behavior on the TUNL task has been analyzed by averaging across the different possible stimulus configurations within a separation condition to create a single summary measurement per animal per separation and delay (i.e., averaging across the different configurations within each row of Fig. 1*B*). When we conducted an equivalent analysis in Datasets 1 and 2 (which tested different separation conditions), we found the usual effect of increasing accuracy with increased separation between matching and nonmatching locations (Dataset 1: β = 0.05, $p < 0.001$; Dataset 2: β = 0.11, $p < 0.001$). However, this summary approach implicitly assumes that different stimulus configurations within a separation condition are otherwise interchangeable with one another, and that the only feature of a sample/nonmatch location pair that is relevant for behavior is the amount of spatial separation between the two locations. We next tested whether this assumption was met in our data by testing for differences in performance across different configurations within a separation condition.

As can be seen in Figure 2, this assumption was not met: instead of different configurations falling on the same temporal decay curve (as we would expect if configurations were interchangeable with one another), there was marked heterogeneity between configurations within each separation condition. Mixed-effects logistic regression analyses revealed that these differences were statistically significant in all three of our datasets: for Dataset 1 (Fig. 2*A*), there was a significant difference between different configurations on accuracy for S0 trials ($\chi^2_{(7)} = 287.26$, $p < 0.001$), S1 trials ($\chi^2_{(5)} = 402.39$, $p < 0.001$), and S2 trials ($\chi^2_{(3)} = 170.14$, $p < 0.001$), although the effect of configuration was not statistically significant for S3 trials ($\chi^2_{(1)} = 1.67$, $p = 0.20$). For Dataset 2 (Fig. 2*B*), there was a significant effect of configuration on accuracy for S1 trials ($\chi^2_{(5)} = 21.18$, $p < 0.001$) and S2 trials ($\chi^2_{(3)} = 23.06$, $p < 0.001$), although again the effect of configuration was not statistically significant for S3 trials ($\chi^2_{(1)} = 0.08$, $p = 0.77$). For Dataset 3 (Fig. 2*C*), which used a smaller number of unique trial configurations, the effect of configuration on accuracy for S1 trials ($\chi^2_{(3)} = 5.79$, $p = 0.12$) was not statistically significant, but there was a significant interaction between configuration and delay ($\chi^2_{(3)} = 19.96$, $p < 0.001$), indicating that the rate at which performance deteriorated as a function of delay differed between the different configurations. These results indicate that, contrary to common assumptions, there are widespread differences in TUNL performance depending on the exact configuration of sample location and nonmatch location that was tested.

Between-configuration differences are consistent with prospective WM coding

We next sought to unravel the source of these between-configuration differences in behavior. In particular, we investigated a striking pattern in which some pairs of Separation-0 stimulus configurations showed markedly different patterns of performance from one another, although they shared the same pair of response options in the choice phase of the trial.

One such pair (from Dataset 1) is presented in Figure 3*A*. When the sample location was at the far left and the correct nonmatch response location was second from the left, animals performed the task well above chance-level performance at zero delay (mixed-effects logistic regression; intercept = 1.26, $p < 0.001$), and this performance level slowly deteriorated with increases in the duration of the retention interval (β = −0.05, $\chi^2_{(1)} = 22.99$, $p < 0.001$). Notably, however, when these locations were reversed (sample location second from left and correct nonmatch response location at far left), animals performed the task significantly below chance performance at zero delay (intercept = −0.49, $p < 0.001$), and performance improved with increases in the duration of the retention interval ($\beta = 0.02$, $\chi^2_{(1)} = 5.99$, $p = 0.01$). This pattern was not restricted to the left side of the testing arena (for a similar pattern, see Extended Data Fig. 3-1).

This striking pattern of results is difficult to reconcile with the standard hypothesis that animals are performing this task using spatial WM (i.e., a retrospective memory code; maintaining a representation in WM of the spatial location of the sample stimulus). Under this standard explanation, the difficulty of Separation-0 trials results from the strong spatial similarity in the two possible response locations in the choice phase locations. Crucially, however, this similarity is symmetric: by definition, the spatial similarity of the far-left location compared with the second-from-left location is identical to the spatial similarity of the second-from-left location compared with the far-left location. Because of this symmetry, it is difficult to explain under the assumption of retrospective WM coding how behavior could be significantly above-chance (and deteriorating with increasing delays) in Figure 3*A* (top) but significantly below-chance (and improving with increasing delays) in Figure 3*A* (bottom).

By contrast, this puzzling pattern of effects is relatively easily explained if we instead assume that animals were using prospective WM coding. Recall that under a prospective WM code, animals would maintain in WM during the retention interval a behavioral intention. This is possible because, depending on the sample location, the animal can often predict with reasonable accuracy the direction in which it should respond on the basis of the sample location that it sees. For instance, if the sample location is at the far left, then the animal should always choose whichever response option is further to the right at the choice phase, independent of separation condition (Fig. 1*B*, left column; Extended Data Fig. 1-1). Similarly, if the sample location is second from the left, then in 75% of trials the correct response location will be whichever choice option is further to the right in the choice phase (Fig. 1*B*, second column from left; Extended Data Fig. 1-1); this means that the TUNL task can be partially solved using prospective WM coding. Then, if the animal is using a prospective WM code, when it observes a sample location second from the left (Fig. 3*A*, bottom), it will encode in WM a prospective intention to choose whichever response option is further right, which will produce below-chance performance in this particular case. Moreover, its performance will also improve with increasing delays for this configuration because the animal's choices will become more random (and therefore more correct in this case, until it reaches chance level) as the prospective WM representation fades. Conversely, because trials with the sample location in the center cannot be solved with prospective WM, these trials may be particularly indicative of retrospective coding in WM.

Animals also show idiosyncratic side biases and distal-response biases

We also investigated how behavior in the TUNL task was influenced by two response biases unrelated to WM: side biases and distal-response biases. Side biases refer to the tendency for individual animals to prefer leftward or rightward response options, and have been observed across species and operant conditioning paradigms (Alber and Strupp, 1996; Miletto Petrazzini et al., 2020). By contrast, distal-response biases[2] are more specific to the TUNL task, and refer to the tendency of mice to prefer response locations closer to the left or right walls of the testing arena over more central response locations (see Kim et al., 2015).

We found evidence for both side biases and distal-response biases in all three datasets (Fig. 3*B*–*D*). For side biases (Fig. 3, middle column), although there was no average preference for leftward versus rightward responses at the group level in any individual dataset (Dataset 1: $t_{(82)} = 1.63$, $p = 0.11$; Dataset 2: $t_{(43)} = 1.80$, $p = 0.08$; Dataset 3: $t_{(89)} = 1.23$, $p = 0.23$), permutation tests revealed statistically significant side biases within the behavior of a majority of individual animals in each dataset. In Dataset 1, 27 of 83 mice (33%) showed a significant leftward bias and 43 of 83 mice (52%) showed a significant rightward bias; in Dataset 2, 8 of 44 mice (18%) showed a significant leftward bias and 15 of 44 mice (34%) showed a significant rightward bias; in Dataset 3, 7 of 36 mice (19%) showed a significant leftward bias and 12 of 36 mice (33%) showed a significant rightward bias. Across all datasets, there was no evidence for an association between the strength of animals' side biases and their accuracy rate on the task (Dataset 1: Pearson $r_{(83)} = -0.08$, $p = 0.49$; Dataset 2: $r_{(44)} = -0.17$, $p = 0.27$; Dataset 3: $r_{(36)} = -0.12$, $p = 0.48$).

Distal-response biases (Fig. 3, right column) were less prominent overall than side biases, but nevertheless accounted for a proportion of group-level and animal-level variance. At a group level, we found a significant preference for distal response options over central response options in both Datasets 1 and 2 (Dataset 1: $t_{(82)} = -7.27$, $p < 0.001$; Dataset 2: $t_{(43)} = -2.55$, $p = 0.01$), though not in Dataset 3 (Dataset 3: $t_{(35)} = -1.56$, $p = 0.12$). Moreover, permutation tests revealed statistically significant distal- or central-response biases within the behavior of a sizeable proportion of individual animals in each dataset. In Dataset 1, 52 of 83 mice (63%) showed a significant distal-response bias and 9 of 83 mice (11%) showed a significant center-response bias; in Dataset 2, 5 of 44 mice (11%) showed a significant distal-response bias and 3 of 44 mice (7%) showed a significant center-response bias; in Dataset 3, 8 of 36 mice (22%) showed a significant distal-response bias and 7 of 36 mice (19%) showed a significant center-response bias. It is noteworthy that, although there was a significant group-level preference for distal response options on average, some individual animals showed a significant preference for the center response options. This speaks to the heterogeneity of behavior in the TUNL task and raises the possibility that individual animals might perform the task using different combinations of WM coding schemes and response biases. In Datasets 1 and 2, there was no evidence for an association between the strength of animals' distal-response biases and their accuracy rate on the task (Dataset 1: Pearson $r_{(83)} = 0.03$, $p = 0.82$; Dataset 2: $r_{(44)} = 0.01$ $p > 0.99$). In Dataset 3, animals that preferred central response options more strongly tended to perform better on the task overall ($r_{(36)} = 0.44$, $p < 0.01$). Given that Dataset 3 tested only S1c trials (i.e., trials in which either the sample location or the nonmatching location was in the center), this suggests that animals that were better able to suppress the instinct to move toward the walls of the testing chamber when the correct response location was in the center might have performed better on the task overall.

Task behavior is best explained by a mixture of WM codes and response biases

The model-agnostic analyses above provide evidence that behavior in the TUNL task was consistent with both retrospective and

[2] For succinctness we here refer to "distal-response biases," even though some mice preferred central response locations. In the nomenclature that we adopt here, such a preference would be represented as a *negative* distal-response bias (i.e., a preference *away from* distal response locations)

prospective WM. In addition, these analyses revealed that individual animals' choice behavior was affected by animal-specific response biases (both side biases and distal-response biases). To dissect the relative contributions of each of these response factors, we next turned to hierarchical Bayesian computational modeling of the data.

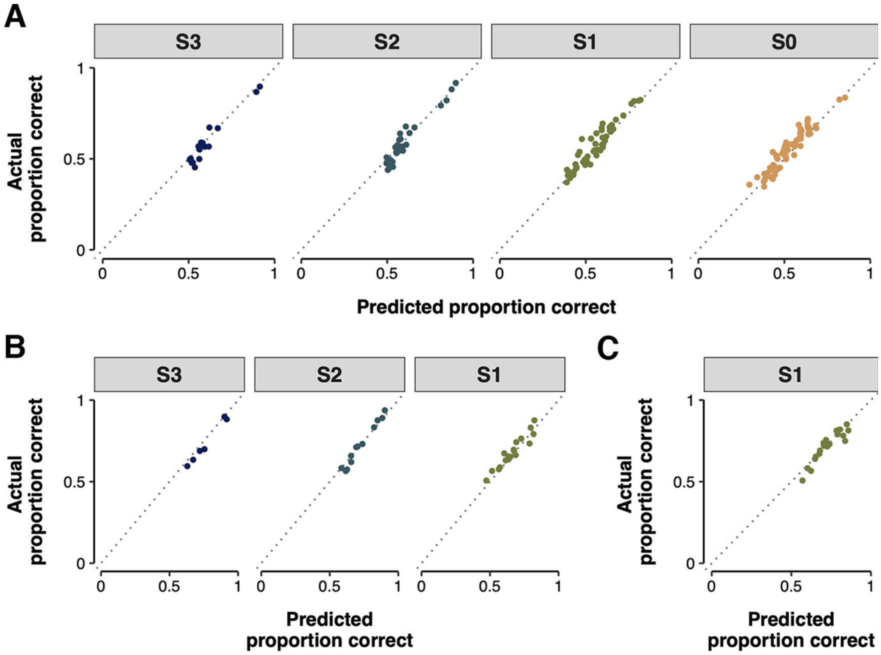

Figure 4. Posterior predictive checks for the best-fitting computational model on each of the three datasets: ***A***, Dataset 1. ***B***, Dataset 2. ***C***, Dataset 3. Subplots represent the correspondence between the actual mean proportion of correct responses across mice (*y* axis) and the model's predicted mean proportion of correct responses across mice (*x* axis), with different separation conditions presented in different plot facets and colors. Each point represents a unique combination of sample location, nonmatch location, and delay, averaged across mice. Diagonal dotted line indicates equality between actual and predicted proportion; in a perfectly calibrated model, every point would fall exactly on this diagonal. The best-fitting computational model provides an excellent fit to data in each of the three independent datasets.

We first compared a suite of models that differed in the set of response factors that were assumed to influence choice behavior (Table 2). Consistent with the results reported in previous sections, we found that the best-fitting model overall was model M9, in which behavior was produced by a mixture of both retrospective and prospective WM, as well as being affected by both side biases and distal-response biases (for parameter estimates, see Extended Data Table 2-3). Model M9 was the best-fitting model in all three datasets, indicating that the observed pattern of results was not consistent to any one animal cohort or experimental design, but rather was displayed consistently across all the datasets we analyzed. The one caveat to this otherwise consistent result comes from Dataset 3: although model M9 was still the best-fitting model overall as measured by the WAIC statistic, model M6 provided a statistically equivalent fit to data when model fit uncertainty was accounted for (i.e., the difference in WAIC values between M6 and M9 in Dataset 3 was smaller than the SE of this difference) (compare Bennett et al., 2021; Weber et al., 2022). Since the difference between model M9 and model M6 is that the former includes prospective WM but the latter does not, this result suggests that prospective WM may have accounted for less variance in behavior in Dataset 3 compared with Datasets 1 and 2. In general, however, posterior predictive checks of model M9 (visualized in Fig. 4) indicated that this model provided a good account of data for all three datasets in absolute terms.

In addition, although model M9 included all four response factors, this does not imply that all four factors necessarily accounted for an equal amount of variance in behavior. To investigate this more quantitatively, we conducted a variance partitioning analysis to estimate the amount of unique variance explained by each of the four response factors in M9 (see Table 3). The results of this analysis indicate substantial heterogeneity across datasets in the proportion of variance accounted for by different factors. In particular, the balance between retrospective and prospective WM differed substantially across datasets: in Dataset 1, prospective WM uniquely accounted for 17.9% of group-level behavioral variance and retrospective WM only 3.1%, whereas in Dataset 3 prospective WM uniquely accounted for only 6.7% of variance and retrospective WM accounted for 22%.

Table 3. Partitioning of unique variance explained within model M9

		Unique variance explained by each response factor[a]			
Dataset #	Overall model R^2 (%)	Spatial/retrospective WM (%)	Prospective WM (%)	Distal-response biases (%)	Side biases (%)
1	88.6	3.1	17.9	32.8	4.6
2	91.1	3.9	4.5	7.5	2.8
3	85.6	22.0	6.7	7.6	1.7

[a] Calculated as change in group-level R^2 when removing a response factor from the model.

Individual animals varied in their mixture of WM codes and response biases

The results presented above provide evidence that, at a group level, animal behavior was best explained as a mixture of WM codes and response biases. Of particular note, there was evidence that mouse behavior in each of the three datasets was produced by both prospective and retrospective WM. There are several different ways that this pattern of data might emerge: first, each individual mouse might have used either prospective or retrospective WM, and each dataset might have consisted of mixtures of these two kinds of mouse in varying proportions. Second, such a result might also arise if individual animals used both prospective and retrospective WM coding. We next sought to determine which of these two possibilities provided the best account of data for individual animals (see Fig. 5).

This analysis revealed three notable features of the data. The first is that, in all three datasets, every mouse used either a prospective WM code, a retrospective WM code, or both. This result is to be expected given that all animals were trained to criterion on the task, but nevertheless confirms that model M9 provided a good fit to the behavior of individual animals as well as of group-level behavior. Second, an overwhelming majority of animals in all three datasets (95% of mice in Dataset 1, 100% of mice in Dataset 2, 83% of mice in Dataset 3) showed evidence of either distal-response biases, side biases, or both. This indicates that these idiosyncratic response biases were a pervasive feature of animal behavior on the TUNL task.

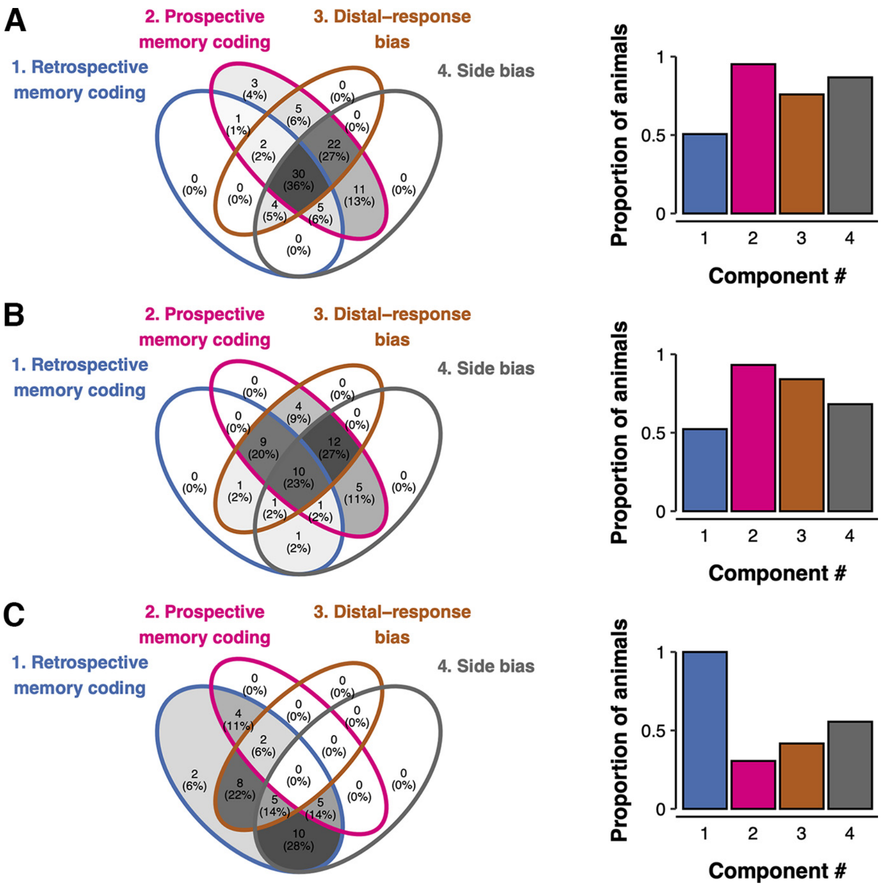

Figure 5. Breakdown of significant model components across mice in each dataset: ***A***, Dataset 1. ***B***, Dataset 2. ***C***, Dataset 3. Venn diagrams represent the breakdown of the numbers of mice who fit into each possible configuration of significant model components (presented as raw numbers and percentage of dataset). Histograms represent the proportion of mice in each dataset that displayed a significant effect of each model component overall (i.e., the sum of the percentages within each colored circle in the Venn diagrams). Results of an equivalent analysis in subsets of trials from Datasets 1 and 2 designed to match the trial types tested in Dataset 3 can be found in Extended Data Figure 5-1.

Third, this analysis revealed that the distribution of prospective and retrospective WM coding was similar across animals in Datasets 1 and 2, but that animals in Dataset 3 showed a qualitatively different pattern of effects. Within Datasets 1 and 2, a large majority of mice showed effects of prospective WM (95% of animals in Dataset 1, 93% of animals in Dataset 2), whereas the proportion of animals using retrospective WM was significantly smaller in these datasets: 51% of animals in Dataset 1 ($\chi^2_{(1)} = 39.51$, $p < 0.001$, two-sample test for equality of proportions) and 52% of animals in Dataset 2 ($\chi^2_{(1)} = 16.56$, $p < 0.001$). In Dataset 3, by contrast, this difference was significant in the opposite direction: only 31% of animals showed an effect of prospective WM, but 100% showed an effect of retrospective WM ($\chi^2_{(1)} = 35.30$, $p < 0.001$). Simultaneous use of prospective and retrospective WM was relatively common within individual animals in all three datasets (although Dataset 3 once again showed somewhat discrepant results). Within Datasets 1 and 2, relatively few mice used only retrospective WM (5% and 7%, respectively), whereas a plurality of mice used either prospective WM only (49% and 47%, respectively) or both retrospective and prospective WM (46% and 45%, respectively). By contrast, in Dataset 3, the majority of animals used only retrospective WM (69%), none used only prospective WM, and 31% used both retrospective and prospective WM.

The most likely explanation for the discrepant results across datasets is that animals were tested on different trial types in the different datasets: specifically, mice in Dataset 3 were only tested in Separation-1 trials involving the central response location (see Table 1), which is a much smaller subset of trial types than were assessed in Datasets 1 and 2. We therefore sought to ensure that the differing model comparison results between datasets were not a statistical artifact of this difference by refitting models to a subset of trials in Datasets 1 and 2 that corresponded to the subset of trials tested in Dataset 3 (i.e., S1c trials only). Our rationale for this analysis is that, if the differences between datasets evident in Figure 5 were solely because of the fact that different separation conditions were tested in different datasets, then we would expect an analysis of S1c trials in Datasets 1 and 2 to produce a very similar response factor distributions to what we observed in Dataset 3. The results of this analysis (see Extended Data Fig. 5-1) indicated that choice behavior on S1c trials within Datasets 1 and 2 was not dominated by prospective WM, as was the case in Dataset 3. We therefore conclude that between-datasets differences in the distributions of response factors are likely to reflect the use of different patterns of response factors by mice in Dataset 3 compared with Datasets 1 and 2. Moreover, this result also suggests that, to measure primarily retrospective WM with the TUNL, it may be necessary to change the entire set of separations that is tested in the probe phase. Our results suggest that performance on specific trial types can be affected by mice's experience with other trial types (particularly when exposed to sessions of mixed trial types), and so it is likely to be insufficient simply to extract S1c trials and analyze those trials in isolation, for instance. Conversely, one alternative hypothesis is suggested by between-dataset differences in the sex of the animals (all male in Dataset 3, compared with a mix of male and female mice in Datasets 1 and 2). Although we consider the effects of testing configuration to be a more likely explanation for between-dataset differences in use of prospective WM, future research could test female mice solely on the S1c configurations from Dataset 3 to rule out sex as an explanation for this phenomenon.

Prospective WM coding may reflect rational allocation of cognitive resources

The results presented in the previous section raise two important questions: first, given that prospective WM is a suboptimal strategy on the TUNL task (as detailed in Extended Data Fig. 1-1), why might so many animals have used prospective memory instead of retrospective WM? Second, why might the distribution of response factors across animals have been so markedly different in Dataset 3 compared with Datasets 1 and 2? These differences might merely reflect unexplained variance within specific animal cohorts or testing facilities; however, another possibility is that the different sets of trial types used in the different studies differed in the degree to which they gave mice an incentive to use prospective versus retrospective WM.

Here we propose that these two questions can both be answered by considering the balance between prospective and retrospective WM in each dataset in terms of animals' adaptive responses to the behavioral and cognitive demands of the specific configurations of trial types that were tested. This is consistent with a resource-rational perspective, in which animals' behavior is understood as reflecting a trade-off between the competing demands of reward maximization and physical/cognitive effort minimization (e.g., Lieder and Griffiths, 2020). We specifically propose that, across all datasets, retrospective WM was more cognitively effortful (i.e., required animals to use more cognitive resources) than prospective WM. In Datasets 1 and 2, mice might therefore have favored prospective WM as the less effortful of the two coding schemes (despite its lower expected reward). By contrast, the greater relative reward to be obtained[3] via retrospective WM in Dataset 3 might have given mice in this dataset an incentive to use retrospective WM despite the greater cognitive demands of this response factor.

There are three different lines of evidence in support of this hypothesis: one theoretical, one empirical, and one based on simulations of model performance. From a theoretical perspective, prospective WM is likely to require fewer neurocognitive resources than retrospective WM in the TUNL task because there are fewer possible distinct items that might be coded in prospective WM (two possible prospective response directions vs five possible retrospective sample locations), therefore requiring less memory allocation. Moreover, forgetting may have been more rapid in retrospective WM because of greater interference between representations of the five different sample locations, compared with the lesser similarity of the two diametrically opposed response directions in prospective WM (Wickelgren, 1965; Bunting, 2006). This would be consistent with Figure 3*A*, in which the effects of prospective WM take ∼15 s to return to chance-level performance.

This resource-rational explanation makes testable predictions for the data. If it was harder for animals to maintain retrospectively coded items in WM than prospectively coded items, we would also expect that information should be lost more rapidly from retrospective WM than from prospective WM. To test whether this was the case, we formulated an additional set of computational models in which the forgetting rate parameter β was allowed to vary between retrospective and prospective WM. The results of this additional set of model comparisons (see Extended Data Table 2-4) provided evidence that, in line with our resource-rational explanation, information was lost more rapidly from retrospective WM than from prospective WM in Dataset 1 (difference in means of group-level β distributions = 0.38; 95% Bayesian HDI: [0.01, 0.86]) and Dataset 2 (mean difference = 3.41; 95% Bayesian HDI: [2.46, 4.53]), although this difference was not credibly different from 0 in Dataset 3 (mean difference = −0.21; 95% Bayesian HDI: [−1.23, 0.41]). We note that the finding that information was lost from prospective WM over time (albeit more slowly than from retrospective WM) is also a point of evidence in favor of interpreting this response factor as memory per se (rather than, for instance, in terms of explicit rehearsal behaviors or postural bias).

Finally, simulated model performance provides further evidence that animals' use of prospective WM may have been rational. If information can be held for longer in prospective WM than in retrospective WM, this may mean that animals can acquire more reward using prospective WM at longer retention intervals. This is illustrated via computational simulations of the different WM coding schemes in Figure 6. In particular, the empirical forgetting rates in the best-fitting model in each dataset reveal the resource-rationality of animals' observed strategy: under the observed WM forgetting rates (i.e., β parameters) in Datasets 1 and 2, there was no substantial reward advantage to be gained by using retrospective WM. In Dataset 3, by contrast, for the observed empirical WM forgetting rate, animals could obtain more reward using retrospective WM than using prospective WM. This resource-rational analysis can therefore explain why animals in Dataset 3 used primarily retrospective WM, even as animals in Datasets 1 and 2 used a mixture of both WM types. The results of this analysis also suggest that mouse WM, and especially retrospective WM, is best assayed at short retention intervals (not >6 s), before the memory trace is entirely forgotten.

Extended Data Figure 6-1 includes further material related to the effect of different trial types on the incentives for animals to use prospective versus retrospective WM, including recommendations on the specific trial types that best isolate retrospective WM in the TUNL task. In short, our simulations suggest that use of trial types at small spatial separations (S0 and S1) promote use of retrospective WM, whereas trial types at larger spatial separations (S2 and S3) will tend to promote use of prospective WM. As such, we suggest that probe-phase trials at small spatial separations should be used if experimenters wish to promote use of spatial WM (as distinct from memory for a prospective behavioral response) in mice completing the TUNL task.

Discussion

By automating large-scale behavioral data collection, touchscreen tasks represent a paradigm shift in the assessment of translational cognitive phenotypes (Bussey et al., 2012; Oomen et al., 2013). A crucial outstanding question, however, concerns the external validity of these tasks: that is, whether touchscreen tasks measure cognitive processes in animals that are equivalent to the cognitive processes that are impaired in human neuropsychiatric disorders (Pound and Ritskes-Hoitinga, 2018). In this project, we used computational modeling to investigate the neurocognitive processes that underlie the touchscreen TUNL task of spatial WM in mice. Our goal was to use the rich data produced by the touchscreen paradigm to dissect TUNL behavior into its cognitive subprocesses, thereby appraising to what extent the TUNL task measures an analog of human spatial WM. To our knowledge, previous work has not considered behavior in the TUNL task through the prism of the prospective/retrospective WM dichotomy. Instead, it is often assumed that TUNL behavior is a reflection of spatial WM, which in this context we interpret as a memory representation of the spatial location of a sample stimulus, that is, a retrospective WM code. Our results suggest that in addition to this spatial/retrospective WM code, mice also made use of a prospective WM code. The best-fitting computational model also included two types of response biases, representing animal-specific preferences for responding at particular spatial locations independent of WM processes.

Across three separate datasets, our computational modeling results consistently showed that behavior was best explained by a

[3] Dataset 3 tested a different subset of trial types to Datasets 1 and 2 (for details, see Table 1). Specifically, Dataset 3 tested solely "S1c" trials, in which prospective working memory cannot produce above-chance performance (see Extended Data Fig. 1-1). The advantage for retrospective working memory is therefore significantly greater in Dataset 3 than in Datasets 1 and 2.

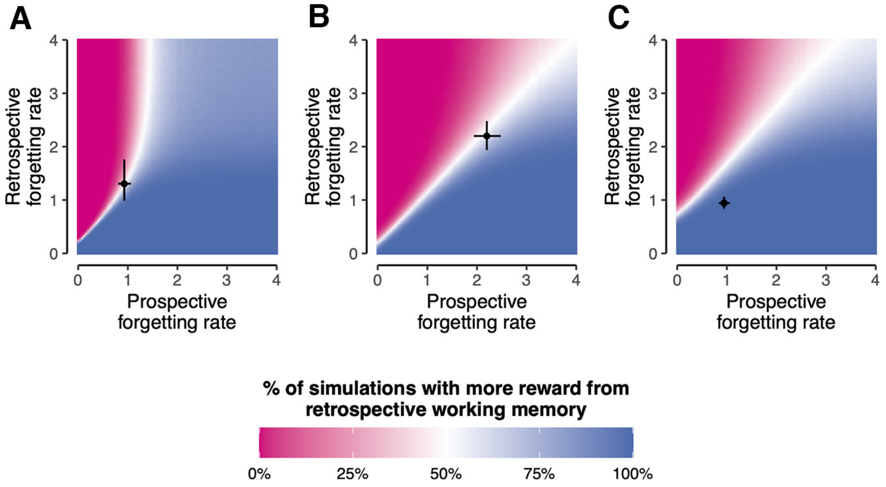

Figure 6. Simulated performance under different forgetting rates for prospective and retrospective WM in the best fitting model for each dataset: ***A***, Trial types and delays as per Dataset 1. ***B***, Trial types and delays as per Dataset 2. ***C***, Trial types and delays as per Dataset 3. Blue (/pink) regions of each heatmap represent parameter regimens under which more reward can be acquired using retrospective (/prospective) WM. White regions represent parameter regimens in which retrospective and prospective WM produce equivalent performance on average. Superior performance for prospective WM typically occurs when the forgetting rate for prospective WM is markedly lower than the forgetting rate for retrospective WM, although comparison of the three panels indicates that the exact region of superiority depends on the trial types that are tested. Points and error bars in each dataset represent the mean estimated forgetting rate for each WM type in each dataset ± the 95% credible interval of the mean. For Datasets 1 and 2, this estimated mean falls within a zone of indifference between prospective and retrospective WM; whereas for Dataset 3, the estimated mean falls within a region in which retrospective WM produces superior performance. Higher forgetting rates correspond to more rapid loss of information from WM. Equivalent plots for simulated experiments testing different separation conditions can be found in Extended Data Figure 6-1.

combination of retrospective WM (i.e., encoding of the spatial location of the prior sample stimulus) and prospective WM (i.e., encoding an intended future behavioral policy). Although this finding stands in contrast to the standard assumption that behavioral strategy in this task solely reflects a purely retrospective WM code, the consistency of model comparison results across datasets indicates that the TUNL reliably assays the same set of four response factors across testing parameters, as well as across different genotypes and experimental manipulations. In our results, the proportion of behavioral variance that was uniquely accounted for by retrospective WM was never >22% in any of our three datasets; we therefore conclude that in mice, performance on the TUNL task does not solely reflect spatial WM. As such, great care is needed in translational neuroscience research using this task, where it is critical to ensure that impairments observed in a given genetic, pharmacological, or acquired model of a psychiatric or neurologic disorder do indeed reflect deficits in retrospective WM (thus capturing clinically relevant symptoms). Our results suggest the importance of careful analysis to distinguish this possibility both from deficits in prospective WM and from changes in the balance between animals' usage of retrospective versus prospective WM codes.

A robust behavioral finding in the TUNL task is that animals' response accuracy reduces with decreasing spatial separation between the sample location and the nonmatching response location. This finding has previously been taken as evidence that animals use retrospective WM to complete the task (e.g., Kim et al., 2015; Sbisa et al., 2017; Gogos et al., 2020). However, our results show that the same qualitative pattern can also be produced by prospective WM, since closer spatial separations also result in greater prospective response ambiguity for an animal using prospective WM (Extended Data Fig. 1-1). Moreover, mice in our datasets showed distinctive behavioral signatures of prospective WM that cannot be explained by retrospective WM, such as significantly below-chance performance on certain S0 trial configurations (Fig. 2*A*). As such, we suggest that prospective WM may also explain other puzzling phenomena that have been observed in the TUNL literature, such as differences in response accuracy between "center" and "noncenter" trials (i.e., worse performance for trials in which the sample location is in the center). This phenomenon was first documented by Kim et al. (2015), who also showed that trials with a central sample stimulus were more sensitive to hippocampal dysfunction that trials with a noncentral sample stimulus. This result is in line with our proposition that these trials rely on retrospective WM, which is in turn more hippocampus-dependent than prospective WM. In line with this hypothesis, in Dataset 3, we found that a task design that maximized the number of "center" trials maximized animals' use of retrospective WM (and decreased their use of prospective WM).

These results have significant implications for our understanding of the neurocomputational processes that subserve WM in mice. TUNL task performance is often interpreted in terms of the operation of a hippocampal pattern separation algorithm (e.g., Talpos et al., 2010; McAllister et al., 2013; Kumar et al., 2015; Kenton et al., 2018). In hippocampal pattern separation, the hippocampus is involved in creating distinct memory representations for stimuli that are otherwise highly similar to one another; these distinct memory representations are what allow similar experiences to be discriminated from one another in memory (see, e.g., Yassa and Stark, 2011). This process has been posited (e.g., McAllister et al., 2013; Oomen et al., 2013) as a mechanism that animals rely on for the small-separation trials of the TUNL task (e.g., S0 and S1 trials), since in these trials the stimuli to be distinguished (the sample and nonmatching locations) are spatially very similar to one another. As such, behavior in small-separation trials has been posited as a "behavioral readout" of pattern separation (Oomen et al., 2013). However, the interpretation of small-separation trials as a readout of hippocampal pattern separation depends on the assumption that animals are using retrospective WM, because spatially similar stimuli will only be similar to one another in WM if what is being encoded in WM is their spatial location (i.e., a retrospective WM code). By contrast, our results suggest that performance at low spatial separations may depend more strongly on prospective WM processes. Since prospective WM has been linked more strongly with prefrontal processing in both rodents and humans (Okuda et al., 1998; Rainer et al., 1999), our results suggest that mouse behavior on the TUNL task may depend more on the PFC than on hippocampal pattern separation at low spatial separations. It is important to note, however, that this conclusion rests on the assumption that mice are being trained on the TUNL according to a standard conditioning protocol (Kim et al., 2015). As we discuss further below, it is possible that the link between prospective WM and behavior on low-separation trials might be weakened if an alternative TUNL conditioning and/or testing protocol were used. More broadly, we note that both the hippocampus and the PFC have been implicated in both human spatial WM and TUNL task behavior (e.g., Talpos et al., 2010; McAllister et al., 2013), suggesting that the TUNL task may provide a good translational model of the neural correlates of spatial WM in general.

The distinction between prospective and retrospective WM has a long history of study in the animal behavior literature (e.g., Cook et al., 1985; Kametani and Kesner, 1989; Kesner, 1989; Rainer et al., 1999). Consistent with the findings that we report here, Kesner (1989) showed that, depending on task demands, rats use a mixture of prospective and retrospective WM codes to solve a radial maze task (Kesner, 1989). Indeed, human participants also adjust their usage of prospective and retrospective WM depending on task demands (Nallan et al., 1991). However, for human spatial memory tasks that, like the TUNL, assay WM for spatial locations or configurations (e.g., spatial WM tasks, object location memory tasks, and positional memory tasks) (see Kessels et al., 2001), humans are thought to primarily encode a representation of the location or configuration of objects in space (Postma et al., 2004). In other words, in human tasks that are conceptually similar to the TUNL, participants are thought primarily to use a retrospective WM code (Kessels et al., 2001; Postma et al., 2004). As such, for the external validity of the TUNL task, it is important to consider how TUNL task design can be optimized to maximize animals' use of retrospective WM (see also recent work on a 6-location version of the TUNL task by Dexter et al., 2022). Our model simulation analyses suggest that it might be possible to promote the use of retrospective WM by increasing the proportion of probe-phase trials that test trial types which can only be solved using retrospective WM. One way of doing this would be to test only S1c trials (i.e., those in which either the sample location or the nonmatching location was in the center): in Dataset 3, which adopted this approach, we found that animals primarily used retrospective WM. More broadly, our simulation analyses (see Extended Data Fig. 6-1) illustrate that testing animals only on S0 trials, or on a combination of S0 and S1 trials, would be expected to have the same effect. By contrast, our results suggest that testing primarily S2 and S3 trials would be expected to increase reliance on prospective WM.

More broadly, it is unclear to what extent the response factor that we have labeled prospective WM is influenced by mediating behavioral strategies (e.g., orienting the body in the direction of the intended behavioral response). Such "rehearsal through overt behavior" has been widely documented in the literature on delayed nonmatch to sample memory tasks (see, e.g., Dudchenko and Sarter, 1992). Nevertheless, the TUNL task includes design features specifically intended (Talpos et al., 2010; Oomen et al., 2013) to reduce the likelihood of such rehearsal behaviors (e.g., requiring a nose-poke at the back of the testing arena before making a response). Consequently, it is an open question to what extent the response factor that we have labeled prospective WM is associated with explicit motor activity or postural biases. Teasing apart this question is an important task for future research, potentially requiring inspection of camera footage of animals performing the TUNL task.

We also found evidence for two distinct types of animal-specific response biases: side biases and distal-response biases. Side biases, which represent animals' stable and idiosyncratic preferences for responding in a leftward or rightward direction, have been consistently observed in the animal behavior literature (e.g., Treviño, 2014; Kuwabara et al., 2020; Broschard et al., 2021). By contrast, there has been less study of distal-response biases, in which individual animals have stable preferences for responding either closer to the walls or closer to the center of the testing chamber. One possibility is that distal-response biases may be related to the preference for proximity to vertical surfaces that is thought to reduce predation risk for small mammals in naturalistic contexts (e.g., Jensen et al., 2003). In line with this interpretation, we observed that mice showed a significant preference for response options closer to the walls of the chamber in two of the three datasets that we analyzed. As such, we speculate that distal-response biases may reflect an anxiety-related phenotype, similar to that measured in the elevated plus maze (Walf and Frye, 2007).

More broadly, we suggest that these findings are best understood within a resource-rational cognitive framework (Lieder and Griffiths, 2020), which assumes that animals consider the cognitive effort costs associated with different behavioral strategies as well as the reward that can be obtained via each strategy. Accordingly, we suggest that mice completing the TUNL task might have preferred to adopt the less effortful approach of using prospective WM, rather than the more effortful approach of retrospective WM, although somewhat more reward could be obtained via retrospective WM. This explanation can account both for the surprisingly high usage rates of prospective WM in a task that was designed to elicit retrospective WM, as well as for the finding that mice appeared to forget information more rapidly from retrospective WM than from prospective WM in our data. Of course, it is important to note that rate of forgetting is only one dimension along which animal WM might differ from human WM. Other dimensions of WM are also pertinent, such as the number of representations that can be maintained. Since the TUNL only requires maintenance of a single item in memory, and WM deficits in disorders, such as schizophrenia, show reduced capacity to hold multiple items in memory (see, e.g., Gold et al., 2010), this is a limitation that future rodent behavioral studies wishing to align rodent models with WM deficits in schizophrenia should seek to rectify. More broadly, it may be useful to develop an exact equivalent of the TUNL task in human participants to truly maximize the comparability of behavior across species.

Several limitations of our computational modeling approach should be noted here. First, the analyses presented here focus only on behavior in mice, but there also exists a version of the TUNL task for rats, using a 3 × 5 grid of possible response locations (Oomen et al., 2013; Sbisa et al., 2017) instead of the 1 × 5 grid (or 1 × 6 grid; see Dexter et al., 2022) tested in mice. It is an empirical question whether our modeling results also extend to rat behavior on the TUNL task, and it is likely that more complex models will be required given the greater complexity of the rat version of the task and the greater cognitive sophistication of rats. Second, we fit models to aggregate data across animals; and as such, our results do not shed light on the trial-by-trial dynamics of WM. One possibility is that animals strategically used prospective WM for some trial types and retrospective WM for other trial types. An alternative possibility is that, for each animal, behavior involved a mixture of prospective and retrospective WM in a consistent way across different trial types. Further within-trial data (e.g., *in vivo* neural recordings from hippocampus or PFC) are required to adjudicate between these two possibilities. Finally, because mice were trained using a standard training protocol in each of the datasets that we analyzed here, our results do not shed light on the extent to which animals' use of prospective WM might be an artifact produced by the training protocol itself. The training phase of the TUNL task typically begins with the easier high-separation trials and only proceeds to the more difficult low-separation trials after animals reach criterion on the earlier trials. Because retrospective WM is only advantageous over prospective WM for low-separation trials, this training protocol may inadvertently have the effect of increasing the salience of the prospective WM strategy. An open question is whether changes in task training, such as those suggested by Dexter et al. (2022), might alter the overall balance between prospective and retrospective WM. Another important open question is the degree to which prospective and retrospective WM is each susceptible to the effects of proactive interference from the correct response in the previous trial (compare Dunnett and Martel, 1990; Dunnett et al., 1990). Understanding this point may prove important, since pharmacological or other interventions that strengthen WM may have the paradoxical effect of worsening animals' performance on subsequent short-delay trials via proactive interference. Dunnett and Martel (1990) suggest extended intertrial intervals as one possible work-around for this issue.

In conclusion, across three distinct datasets, our computational modeling results provide evidence that the behavior of mice on the TUNL task is more multifaceted than has often been appreciated. Specifically, we found that retrospective WM, which has often been assumed to be the dominant factor underlying TUNL performance, only accounted for a portion of the variance in the data. Indeed, in two of the three datasets that we studied, prospective WM was a more significant factor in mouse behavior than retrospective WM. Of course, this result does not entail that the TUNL task lacks translational validity as an assay of spatial WM; rather, our results suggest that retrospective (i.e., spatial) WM is an important component of behavior on the task, and that it is incumbent on the task design to be optimized for this purpose to minimize other cognitive strategies. Specifically, our results suggest a number of factors that might be adjusted in future research using the TUNL task to maximize its external validity as a measure of spatial WM: in particular, maximizing the number of S1-center trials. Where this is not possible, our results suggest that computational modeling provides a tractable way for isolating variance in behavior that is uniquely associated with spatial WM. In addition, future research could also analyze additional dependent variables that can be readily extracted from touchscreen behavioral tasks, such as response latencies. Each of these approaches may prove beneficial in optimizing the utility of the TUNL task as a translational assay in rodent models of neurodevelopmental and psychiatric disorders.

References

Alber SA, Strupp BJ (1996) An in-depth analysis of lead effects in a delayed spatial alternation task: assessment of mnemonic effects, side bias, and proactive interference. Neurotoxicol Teratol 18:3–15.

Aultman JM, Moghaddam B (2001) Distinct contributions of glutamate and dopamine receptors to temporal aspects of rodent working memory using a clinically relevant task. Psychopharmacology (Berl) 153:353–364.

Baddeley A (2010) Working memory. Curr Biol 20:R136–R140.

Barch DM, Berman MG, Engle R, Jones JH, Jonides J, Macdonald A, Nee DE, Redick TS, Sponheim SR (2009) CNTRICS final task selection: working memory. Schizophr Bull 35:136–152.

Barch DM, Moore H, Nee DE, Manoach DS, Luck SJ (2012) CNTRICS imaging biomarkers selection: working memory. Schizophr Bull 38:43–52.

Barr DJ, Levy R, Scheepers C, Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: keep it maximal. J Mem Lang 68:255–278.

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Soft 67:1–48.

Bennett D, Radulescu A, Zorowitz S, Felso V, Niv Y (2021) Affect-congruent attention modulates generalized reward expectations. PsyArXiv.

Bloch S, Froc C, Pontiggia A, Yamamoto K (2019) Existence of working memory in teleosts: establishment of the delayed matching-to-sample task in adult zebrafish. Behav Brain Res 370:111924.

Broschard MB, Kim J, Love BC, Freeman JH (2021) Category learning in rodents using touchscreen-based tasks. Genes Brain Behav 20:e12665.

Bunting M (2006) Proactive interference and item similarity in working memory. J Exp Psychol Learn Mem Cogn 32:183–196.

Bussey TJ, Holmes A, Lyon L, Mar AC, McAllister KA, Nithianantharajah J, Oomen CA, Saksida LM (2012) New translational assays for preclinical modelling of cognition in schizophrenia: the touchscreen testing method for mice and rats. Neuropharmacology 62:1191–1203.

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A (2017) Stan: a probabilistic programming language. J Stat Softw 76:1.

Carruthers P (2013) Evolution of working memory. Proc Natl Acad Sci USA 110:10371–10378.

Castner SA, Goldman-Rakic PS, Williams GV (2004) Animal models of working memory: insights for targeting cognitive dysfunction in schizophrenia. Psychopharmacology 174:111–125.

Chudasama Y, Muir JL (1997) A behavioural analysis of the delayed non-matching to position task: the effects of scopolamine, lesions of the fornix and of the prelimbic region on mediating behaviours by rats. Psychopharmacology (Berl) 134:73–82.

Cook RG, Brown MF, Riley DA (1985) Flexible memory processing by rats: use of prospective and retrospective information in the radial maze. J Exp Psychol Anim Behav Process 11:453–469.

Cowan N (2014) Working memory underpins cognitive development, learning, and education. Educ Psychol Rev 26:197–223.

Della Sala S, Gray C, Baddeley A, Allamano N, Wilson L (1999) Pattern span: a tool for unwelding visuo–spatial memory. Neuropsychologia 37:1189–1199.

Dexter TD, Palmer D, Hashad AM, Saksida LM, Bussey TJ (2022) Decision making in mice during an optimized touchscreen spatial working memory task sensitive to medial prefrontal cortex inactivation and NMDA receptor hypofunction. Front Neurosci 16:905736.

Donkin C, Nosofsky RM (2012) A power-law model of psychological memory strength in short- and long-term recognition. Psychol Sci 23:625–634.

Dudchenko PA (2004) An overview of the tasks used to test working memory in rodents. Neuroscience & Biobehavioral Reviews 28:699–709.

Dudchenko P, Sarter M (1992) Behavioral microanalysis of spatial delayed alternation performance: rehearsal through overt behavior, and effects of scopolamine and chlordiazepoxide. Psychopharmacology (Berl) 107:263–270.

Dudchenko P, Talpos J, Young J, Baxter MG (2013) Animal models of working memory: a review of tasks that might be used in screening drug treatments for the memory impairments found in schizophrenia. Neurosci Biobehav Rev 37:2111–2124.

Duff SJ, Hampson E (2001) A sex difference on a novel spatial working memory task in humans. Brain Cogn 47:470–493.

Dunnett SB, Martel FL (1990) Proactive interference effects on short-term memory in rats: I. Basic parameters and drug effects. Behav Neurosci 104:655–665.

Dunnett SB, Martel FL, Iversen SD (1990) Proactive interference effects on short-term memory in rats: II. Effects in young and aged rats. Behav Neurosci 104:666–670.

Ferbinteanu J, Shapiro ML (2003) Prospective and retrospective memory coding in the hippocampus. Neuron 40:1227–1239.

Fu S, Czajkowski N, Rund BR, Torgalsbøen AK (2017) The relationship between level of cognitive impairments and functional outcome trajectories in first-episode schizophrenia. Schizophr Res 190:144–149.

Giurfa M, Zhang S, Jenett A, Menzel R, Srinivasan MV (2001) The concepts of 'sameness' and 'difference' in an insect. Nature 410:930–933.

Gogos A, Sbisa A, Witkamp D, Buuse M (2020) Sex differences in the effect of maternal immune activation on cognitive and psychosis-like behaviour in Long Evans rats. Eur J Neurosci 52:2614–2626.

Gold JM, Hahn B, Zhang WW, Robinson BM, Kappenman ES, Beck VM, Luck SJ (2010) Reduced capacity but spared precision and maintenance of working memory representations in schizophrenia. Arch Gen Psychiatry 67:570–577.

Gold JM, Barch DM, Feuerstahler LM, Carter CS, MacDonald AW, Ragland JD, Silverstein SM, Strauss ME, Luck SJ (2019) Working memory impairment across psychotic disorders. Schizophr Bull 45:804–812.

Gold JM, Luck SJ (2022) Working memory in people with schizophrenia. In: Cognitive functioning in schizophrenia: leveraging the RDoC framework (Barch D, Young J, eds), pp 137–152. New York: Springer.

Hopkins RO, Kesner RP, Goldstein M (1995) Item and order recognition memory in subjects with hypoxic brain injury. Brain and Cognition 27:180–201.

Jensen SP, Gray SJ, Hurst JL (2003) How does habitat structure affect activity and use of space among house mice? Anim Behav 66:239–250.

Kametani H, Kesner P (1989) Retrospective and prospective coding of information: dissociation of parietal cortex and hippocampal formation. Behav Neurosci 103:84–89.

Kenton JA, Castillo R, Holmes A, Brigman JL (2018) Cortico-hippocampal GluN2B is essential for efficient visual-spatial discrimination learning in a touchscreen paradigm. Neurobiol Learn Mem 156:60–67.

Kesner RP (1989) Retrospective and prospective coding of information: role of the medial prefrontal cortex. Exp Brain Res 74:163–167.

Kessels RP, de Haan EH, Kappelle LJ, Postma A (2001) Varieties of human spatial memory: a meta-analysis on the effects of hippocampal lesions. Brain Research Reviews 35:295–303.

Kim CH, Romberg C, Hvoslef-Eide M, Oomen CA, Mar AC, Heath CJ, Berthiaume AA, Bussey TJ, Saksida LM (2015) Trial-unique, delayed nonmatching-to-location (TUNL) touchscreen testing for mice: sensitivity to dorsal hippocampal dysfunction. Psychopharmacology (Berl) 232:3935–3945.

Kofler MJ, Sarver DE, Harmon SL, Moltisanti A, Aduen PA, Soto EF, Ferretti N (2018) Working memory and organizational skills problems in ADHD. J Child Psychol Psychiatry 59:57–67.

Kruschke JK (1992) ALCOVE: an exemplar-based connectionist model of category learning. Psychol Rev 99:22–44.

Kumar G, Olley J, Steckler T, Talpos J (2015) Dissociable effects of NR2A and NR2B NMDA receptor antagonism on cognitive flexibility but not pattern separation. Psychopharmacology (Berl) 232:3991–4003.

Kuwabara M, Kang N, Holy TE, Padoa-Schioppa C (2020) Neural mechanisms of economic choices in mice. Elife 9:e49669.

Lieder F, Griffiths TL (2020) Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. Behav Brain Sci 43:e1.

Lind J, Enquist M, Ghirlanda S (2015) Animal memory: a review of delayed matching-to-sample data. Behav Processes 117:52–58.

Logie RH (2014) Visuo-spatial working memory. New York: Psychology.

McAllister KA, Saksida LM, Bussey TJ (2013) Dissociation between memory retention across a delay and pattern separation following medial prefrontal cortex lesions in the touchscreen TUNL task. Neurobiol Learn Mem 101:120–126.

Meteyard L, Davies RA (2020) Best practice guidance for linear mixed-effects models in psychological science. J Mem Lang 112:104092.

Miletto Petrazzini ME, Pecunioso A, Dadda M, Agrillo C (2020) Does brain lateralization affect the performance in binary choice tasks? A study in the animal model *Danio rerio*. Symmetry 12:1294.

Miller EK, Erickson CA, Desimone R (1996) Neural mechanisms of visual working memory in prefrontal cortex of the macacque. J Neurosci 16:5154–5167.

Nakamura JP, Gillespie B, Gibbons A, Jaehne EJ, Du X, Chan A, Schroeder A, van den Buuse M, Sundram S, Hill RA (2021) Maternal immune activation targeted to a window of parvalbumin interneuron development improves spatial working memory: implications for autism. Brain Behav Immun 91:339–349.

Nallan GB, Kennedy K, Kennedy K (1991) Retrospective and prospective memory coding in humans. Psychol Rec 41:79–86.

Nilsson SR, Celada P, Fejgin K, Thelin J, Nielsen J, Santana N, Heath CJ, Larsen PH, Nielsen V, Kent BA, Saksida LM, Stensbøl TB, Robbins TW, Bastlund JF, Bussey TJ, Artigas F, Didriksen M (2016) A mouse model of the 15q13.3 microdeletion syndrome shows prefrontal neurophysiological dysfunctions and attentional impairment. Psychopharmacology (Berl) 233:2151–2163.

Nosofsky RM (1986) Attention, similarity, and the identification-categorization relationship. J Exp Psychol Gen 115:39–61.

Nunn JA, Graydon FJX, Polkey CE, Morris RG (1999) Differential spatial memory impairment after right temporal lobectomy demonstrated using temporal titration. Brain 122:47–59.

Oberauer K, Lin HY (2017) An interference model of visual working memory. Psychol Rev 124:21–59.

Okuda J, Fujii T, Yamadori A, Kawashima R, Tsukiura T, Fukatsu R, Suzuki K, Ito M, Fukuda H (1998) Participation of the prefrontal cortices in prospective memory: evidence from a PET study in humans. Neurosci Lett 253:127–130.

Oomen CA, Hvoslef-Eide M, Heath CJ, Mar AC, Horner AE, Bussey TJ, Saksida LM (2013) The touchscreen operant platform for testing working memory and pattern separation in rats and mice. Nat Protoc 8:2006–2021.

Park S, Holzman PS (1992) Schizophrenics show spatial working memory deficits. Arch Gen Psychiatry 49:975–982.

Postma A, Kappelle RP, van Asselen M (2004) The neuropsychology of object-location memory. In: Human Spatial Memory (pp. 163–180. Psychology Press.

Pound P, Ritskes-Hoitinga M (2018) Is it possible to overcome issues of external validity in preclinical animal research? Why most animal models are bound to fail. J Transl Med 16:304.

Rainer G, Rao SC, Miller EK (1999) Prospective coding for objects in primate prefrontal cortex. J Neurosci 19:5493–5505.

Roberts WA (1972) Short-term memory in the pigeon: effects of repetition and spacing. J Exp Psychol 94:74–83.

Roberts WA, Santi A (2017) The comparative study of working memory. In: APA handbook of comparative psychology: perception, learning, and cognition (Call J, Burghardt GM, Pepperberg IM, Snowdon CT, Zentall T, eds.), pp 203–225. Washington, DC: American Psychological Association.

Roitblat HL (1982) The meaning of representation in animal memory. Behav Brain Sci 5:353–372.

Sbisa AM, Gogos A, van den Buuse M (2017) Spatial working memory in the touchscreen operant platform is disrupted in female rats by ovariectomy but not estrous cycle. Neurobiol Learn Mem 144:147–154.

Shepard RN (1957) Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. Psychometrika 22:325–345.

Shepard RN (1987) Toward a universal law of generalization for psychological science. Science 237:1317–1323.

Sokolenko E, Nithianantharajah J, Jones NC (2020) MK-801 impairs working memory on the Trial-Unique Nonmatch-to-Location test in mice, but this is not exclusively mediated by NMDA receptors on PV^{+} interneurons or forebrain pyramidal cells. Neuropharmacology 171:108103.

Spellman T, Rigotti M, Ahmari SE, Fusi S, Gogos JA, Gordon JA (2015) Hippocampal–prefrontal input supports spatial encoding in working memory. Nature 522:309–314.

Steele SD, Minshew NJ, Luna B, Sweeney JA (2007) Spatial working memory deficits in autism. J Autism Dev Disord 37:605–612.

Talpos JC, McTighe SM, Dias R, Saksida LM, Bussey TJ (2010) Trial-unique, delayed nonmatching-to-location (TUNL): a novel, highly hippocampus-dependent automated touchscreen test of location memory and pattern separation. Neurobiol Learn Mem 94:341–352.

Treviño M (2014) Stimulus similarity determines the prevalence of behavioral laterality in a visual discrimination task for mice. Sci Rep 4:7569.

Troyb E, Rosenthal M, Eigsti IM, Kelley E, Tyson K, Orinstein A, Barton M, Fein D (2014) Executive functioning in individuals with a history of ASDs who have achieved optimal outcomes. Child Neuropsychol 20:378–397.

Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 27:1413–1432.

Walf AA, Frye CA (2007) The use of the elevated plus maze as an assay of anxiety-related behavior in rodents. Nat Protoc 2:322–328.

Watanabe S (2013) A widely applicable Bayesian information criterion. J Mach Learn Res 14:867–897.

Weber ID, Zorowitz S, Niv Y, Bennett D (2022) The effects of induced positive and negative affect on Pavlovian-instrumental interactions. Cogn Emot 36:1343–1360.

Wickelgren WA (1965) Acoustic similarity and retroactive interference in short-term memory. J Verbal Learn Verbal Behav 4:53–61.

Wickelgren WA (1974) Single-trace fragility theory of memory dynamics. Mem Cognit 2:775–780.

Wixted JT, Carpenter SK (2007) The Wickelgren power law and the Ebbinghaus savings function. Psychol Sci 18:133–134.

Yassa MA, Stark CE (2011) Pattern separation in the hippocampus. Trends Neurosci 34:515–525.

Zeleznikow-Johnston A, Burrows EL, Renoir T, Hannan AJ (2017) Environmental enrichment enhances cognitive flexibility in C57BL/6 mice on a touchscreen reversal learning task. Neuropharmacology 117:219–226.

Zimmer H (2008) Visual and spatial working memory: from boxes to networks. Neurosci Biobehav Rev 32:1373–1395.