Mice dynamically adapt to opponents in competitive multi-player games

Jeffrey C. Erlich[1*], Mehul Rastogi[1], Qianbo Yin[1], Ivana Orsolic[1], Bruno F. Cruz[2,3], André Almeida[2], Chunyu A. Duan[1*]

[1]Sainsbury Wellcome Centre for Neural Circuits and Behavior, UCL, London, W1T 4JG, UK.
[2]NeuroGEARS Ltd., UK.
[3]Neural Dynamics, Allen Institute, 615 Westlake Ave North, Seattle, 98109, WA, USA.

*Corresponding authors. E-mails: j.erlich@ucl.ac.uk; c.duan@ucl.ac.uk;

Abstract

Competing for resources in dynamic social environments is fundamental for survival, and requires continuous monitoring of both 'self' and 'others' to guide effective choices. Yet our understanding of value-based decision-making comes primarily from studying individuals in isolation, leaving open fundamental questions about how animals adapt their strategies during social competition. Here, we developed an ethologically relevant multi-player game, in which freely-moving mice make value-based decisions in a competitive spatial foraging task. We found that mice integrate real-time spatial information about 'self' and the opponent to flexibly shift their preference towards safer, low-payout options when appropriate. Analyses of mice and reinforcement learning agents reveal that these behavioural adaptations cannot be explained by simple reward learning, but are instead consistent with optimal decision strategies guided by opponent features. Using a dynamical model of neural activity, we found that in addition to opponent effects, decisions under competition were also noisier and more sensitive to initial conditions, generating testable predictions for neural recordings and perturbations. Together, this work reveals a fundamental mechanism for competitive foraging, and proposes novel quantitative frameworks towards understanding value-based decision-making in a fast-changing social environment.

Keywords: social competition, rodent decision-making, multi-agent, reinforcement learning, attractor dynamics

Introduction and Results

Our understanding of decision-making is largely built on studies of single agents making choices in well-controlled environments [1–6]. We have detailed models of how specific factors, such as sensory evidence [7–9] or costs and benefits [10–14], parametrically influence decision-making. Real life decisions, however, are rarely made by a single agent in isolation: the outcomes of decisions often depend on the choices of other agents in the environment as well as our own actions [15]. Classic games, like rock-paper-scissors, have provided a quantitative approach for studying player strategies [16–23], but the turn-based nature of these games differ from real-life decision scenarios with continuous interactions. On the other hand, studies that have characterized continuous interactions focus on innate behaviors such as prey capture [24, 25] and courtship [26–28]. For understanding the biological basis of multi-agent decision-making, what is missing are tractable models of games, like tennis or football [29, 30], where players need to consider the moment-by-moment actions of other agents in order to make rapid and flexible value-guided decisions.

To study how animals integrate ongoing information of 'self' and 'other' to guide choices, we developed the 'Octagon', a novel area where freely-moving mice harvest reward from eight possible reward patches in a solo or a competitive social context (Fig. 1a,b, Fig. S1, Methods). On each trial, floor-projected visual stimuli (grating or light triangular slices) were presented to indicate the location and value of two active reward patches, each associated with either a high or a low reward (Fig. 1c). Mice indicated their choice by nose-poking at one of the available patches, which triggered the corresponding reward (Fig. 1d, Methods). In the social context, only the first response of a trial was rewarded. The locations of active reward patches changed on a trial-by-trial basis, and the visual stimuli that mapped to the high versus low reward options were counter-balanced across mice.

The Octagon arena combines the embodied nature of spatial mazes [31–35] and the flexibility of virtual reality paradigms [36, 37], providing an ethological yet quantitative approach for studying strategic interactions between multiple animals. This design has several advantages. First, it allows animals to naturally and continuously interact with the environment and with each other during and between decisions. The resulting behavioural variance allows us to parametrically explore how factors such as proximity to reward patches influence animals' foraging decisions. Second, the trial-based structure of the task constrains the timing of animals' patch selection decisions and encourages concurrent decision-making. Third, the ability to flexibly vary the locations of visually cued reward patches across trials dissociates patch quality from patch location, which in turn helps disentangle non-spatial foraging decisions from spatial strategies. Finally, the dynamic nature of the task requires animals to actively survey their surroundings on each trial, including the ongoing behavior of their opponent.

Mice flexibly discriminate patch values in a dynamic foraging environment

Before experiencing social competition in the Octagon, mice were first trained to associate different visual cues with distinct reward sizes in the solo context. Mice learned to discriminate high versus low reward patches over a few weeks (n = 8 mice; Fig. 1e); and their post-learning preference for the high reward patch generalized across patch locations ($82.2 \pm 0.7\%, n = 8$ locations; Fig. 1h, Supplemental Video 1). The sensory-value map was counter-balanced across animals ensuring that the preference for high was driven by value, not sensory features (Fig. S2a). Preferences were also apparent from response times (RT) on 'forced' trials (Methods). When animals had to choose between two low-reward patches, they were significantly slower compared to when choosing between two high-reward patches ($t(7) = 4.29, P = .0036$; paired t-test; Fig. 1f). On a session-by-session basis, this RT difference across forced trial-types was positively correlated with the probability of choosing the high patch on choice trials ($r = .46, P = 1.1 \times 10^{-13}, n = 235$ sessions; Fig. 1g).

In a static environment where the high and low reward patches have fixed locations, animals can form simple spatial strategies such as always staying close to the high patch. In contrast, the dynamic nature of the Octagon encourages active survey of the entire arena and higher-order strategies dissociated from specific patch locations. We tracked animals' continuous positions using Bonsai (Fig. S1c,d; Methods), and quantified how they prepared for each trial, and how their ongoing behaviours affect decision preferences. As expected, mice's estimated positions at the time of response were restricted to one of the eight reward patches (black, Fig. 1i). At stimulus onset, their start positions spanned the entire arena, with a tendency to cluster around the center of the Octagon (gray, Fig. 1i). The strategy of returning to the center in preparation for the new trial emerged without any explicit shaping; and against rodents' innate preference to stay close to walls when exploring open spaces [38]. As a control, we trained a group of mice to associate the grating and light stimuli with equal reward size ('equal' mice; Fig. S3a,b), while keeping all the other training parameters consistent across groups. Unlike the high/low discrimination mice, the 'equal' mice did not develop a strategy of returning

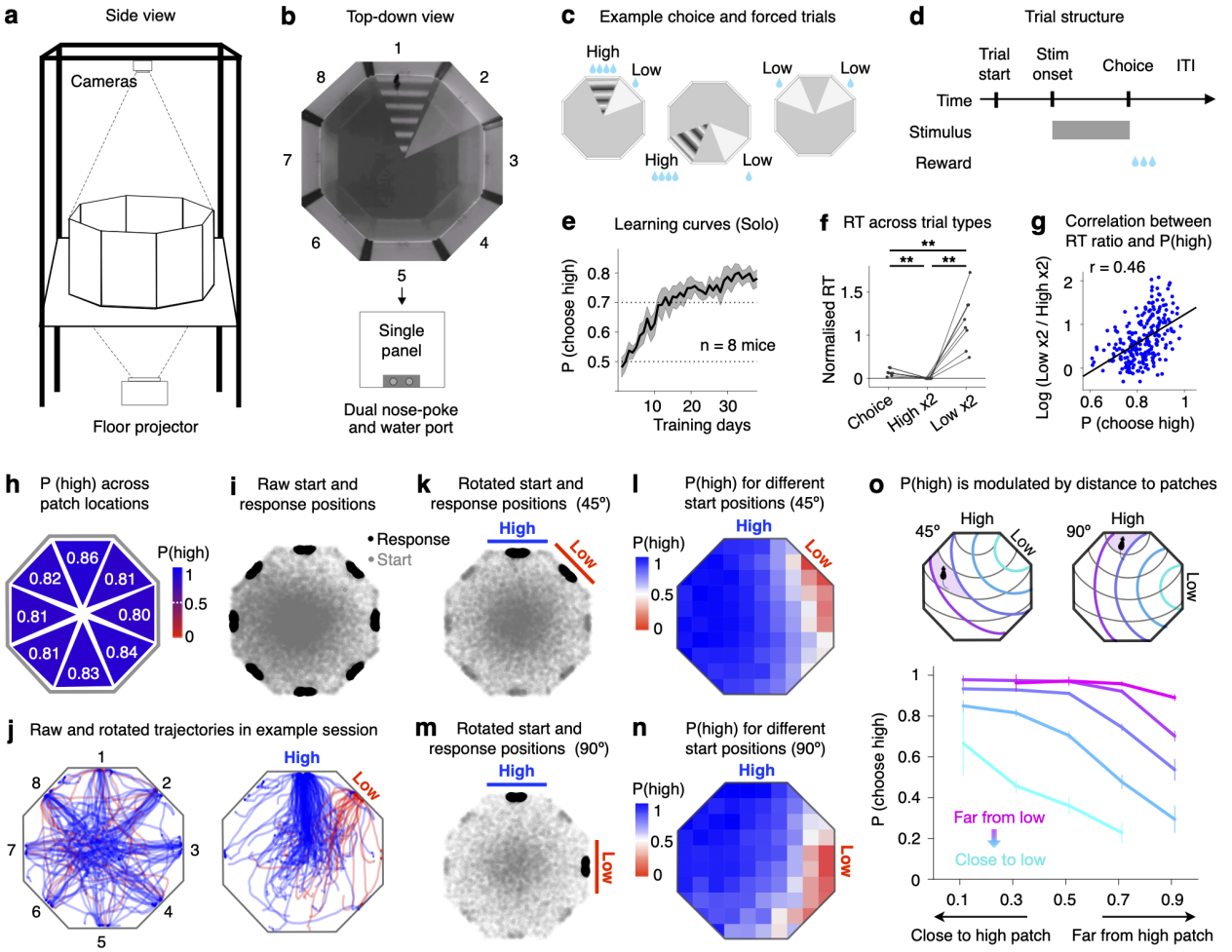

Fig. 1: Mice discriminate and generalize patch values in a novel Octagon arena. **a,b**, Side and top-down view of the Octagon arena. Top cameras track animals' movement; bottom projector displays visual stimuli onto the Octagon floor. Each of the 8 wall panels has nose pokes and water ports for registering choices and delivering reward. **c**, Example floor-projected visual stimuli that cue the location and value of active reward patches on each trial. **d**, Event timing within a trial. **e**, Solo learning curves of 8 experimental mice, mean ± s.e.m. **f**, Log ratio of response times (RT) compared to RTs on high-high forced trials, n = 8 mice. **, $P < .01$, paired bootstrapped tests. **g**, Session-wise correlation between forced trial RT log-ratio and choice trial performance, n = 235 sessions. r, Pearson's correlation coefficient. **h**, Post-learning preference for the high patch across all 8 wall positions, including trials from 235 sessions and 8 mice. **i**, Tracked positions at stimulus onset (gray) and at the time of response (black) for all post-learning solo sessions. **j**, Left, mouse movement trajectories from stimulus onset to response for all choice trials (blue, chose high; red, chose low) in an example session. Right, same data but rotated and reflected for visualization purpose (Methods). **k**, Similar to **i**, but rotated for visualization purpose, only for trials where the high and low reward patches were separated by 45 degrees. **l**, Performance across varied start positions in the Octagon, including all trials shown in **k**. **m,n**, Similar to **k,l**, but for trials where the high and low reward patches were separated by 90 degrees. **o**, Top, schematic illustrating how distances to the high and low patches were binned for different trial types. Bottom, the probability of choosing the high patch as a function of varied distances to the high patch (x axis) and to the low patch (colored lines), error bars are binomial confidence intervals across trials.

to the center at stimulus onset (Fig. S3c,d). We calculated the normalized distance $(0-1)$ between each start position and the arena center for both groups of mice and found that the high/low mice were significantly closer to the arena center (mean distance = 0.56) at stimulus onset compared to the 'equal' mice (mean distance = 0.78, AUC value = 0.73, $P = 2 \times 10^{-4}$, permutation test), and this comparison was individually significant for all high/low mice despite variance across mice (Fig. S2a).

We then used animals' varied start positions to probe how decision preference was modulated by distance to goals. For visualization, we rotated the trajectories of individual trials to align the locations of high and

low reward patches (Fig. 1j,k), and plotted the probability of choosing the high patch as a function of relative start positions in the Octagon. This revealed a spatial gradient of decision preference where mice mostly chose the high patch except when they were much closer to the low patch (Fig. 1l). This trend generalized across mice (Fig. S2a), and was also observed in trials where the high and low patches were further away from each other (Fig. 1m,n). We quantified this result using a generalized linear mixed effects model (GLMM), and found that animals' general preference for the high patch was negatively correlated with distance to the high patch ($\beta_{dist2high} = -2.53, P = 8.65 \times 10^{-108}$) and positively correlated with distance to the low patch ($\beta_{dist2low} = 3.41, P = 1.76 \times 10^{-269}$; Model 2 in Fig. S2c, Fig. 1o, Table S2). Fitting the model to individual mice yielded similar effects per animal (Fig. S2d). Cross-validated model comparison revealed that both distance factors as well as their interaction were important for predicting behavior (Fig. S2c,e). Overall, mice had a consistent preference for the high reward patch that generalized across patch locations and was modulated by the relative distances to the patches. In comparison, the choice preference in the control group was dominated by proximity, since both patches were associated with equal values (Fig. S3e-g).

To test whether returning to the arena center is an advantageous preparation strategy for dynamic foraging in the Octagon, we simulated the average travel distance towards randomly-appearing patches from all start positions in the arena (Methods). Our simulations show that starting trials at the center results in the shortest average travel distance if the goal is to choose a specific patch that randomly appears in the environment (e.g., choose high irrespective of patch locations; Fig. S4a). On the other hand, if the goal is to choose the closest of two randomly appearing patches, as shown in the 'equal' mice, the optimal starting position shifts away from the centre (Fig. S4b,c), as this increases the probability of being close to one of the two patches by chance. Strikingly, the distinct start position patterns observed in the high/low discrimination mice versus in the 'equal' mice match the optimal solutions for their corresponding goals, suggesting that mice have learned policies that consider distance or time costs in addition to reward.

Mice integrate self and opponent information to guide choices under competition

After individual mice learned to discriminate patch values in the Octagon, we introduced competitive social sessions (Fig. 2a, Supplemental Video 2), interleaved with solo control sessions. In the social context, only the first mouse to report its choice was rewarded on each trial (Methods). In other words, the slower-to-respond mouse received no reward. As a result, the effective reward probability and expected values of the two reward options dynamically changed on a trial-by-trial basis, modulated by the ongoing behaviors of 'self' and 'other'. How would mice shift their behavior in this competitive context? One possibility was that mice could ignore their opponent and behave similarly across the social and solo contexts (Fig. 2b). Alternatively, mice could race to the closest patch on most trials and lose their value-based preference. It is also possible that mice could slowly adapt to the changing reward statistics in the social session through reinforcement learning, without discriminating if the reward changes were due to their opponent's actions. Finally, mice could integrate ongoing information of self and other to flexibly guide decisions on each trial, resulting in fast changing behavioral patterns across contexts and across trials.

To test these hypotheses, we compared animals' decision preference in solo versus social contexts, and found a reliable decrease in their preference for the high reward patch under social competition ($10.46 \pm 0.73\%$ decrease, mean $\pm$ s.e.m., $P = 7.20 \times 10^{-17}$; paired t-test across 38 session pairs; Fig. 2c). Despite this shift towards the low-payout option when competing with a conspecific, mice still preferred the high reward patch significantly above chance ($P = 2.84 \times 10^{-23}$), arguing against the first two hypotheses: ignoring the opponent or a proximity-dominated race strategy under competition. If the behavioral changes observed in social sessions resulted from non-social learning over many trials, we might expect larger changes in the second half of the social session compared to the first half; and residual learning effects carried over to next day's solo performance. However, we found no consistent difference between preference on the first versus second half of the social sessions (first half minus second half $= -0.26 \pm 1.02\%, P = 0.80$; Fig. S5d); and the solo performance 1 day after the social session recovered to pre-social solo preference (solo pre minus solo post $= 0.58 \pm 0.70\%, P = 0.41$; Fig. 2c), arguing against a *slow* non-social learning process. Therefore, mice seemed to shift their decision preference flexibly across the solo and social contexts. One potential explanation for the observed decrease in value-based preference is a speed-accuracy trade-off [39]. In other words, maybe the mice seemed less precise (in choosing the high-reward patch) because in the social context, we only analyzed the faster, winning responses. To test that, we generated RT-corrected solo control trials to simulate selection of the faster choice without social competition (Fig. S5a,b). Compared to RT-corrected solo controls, there was still a significant decrease

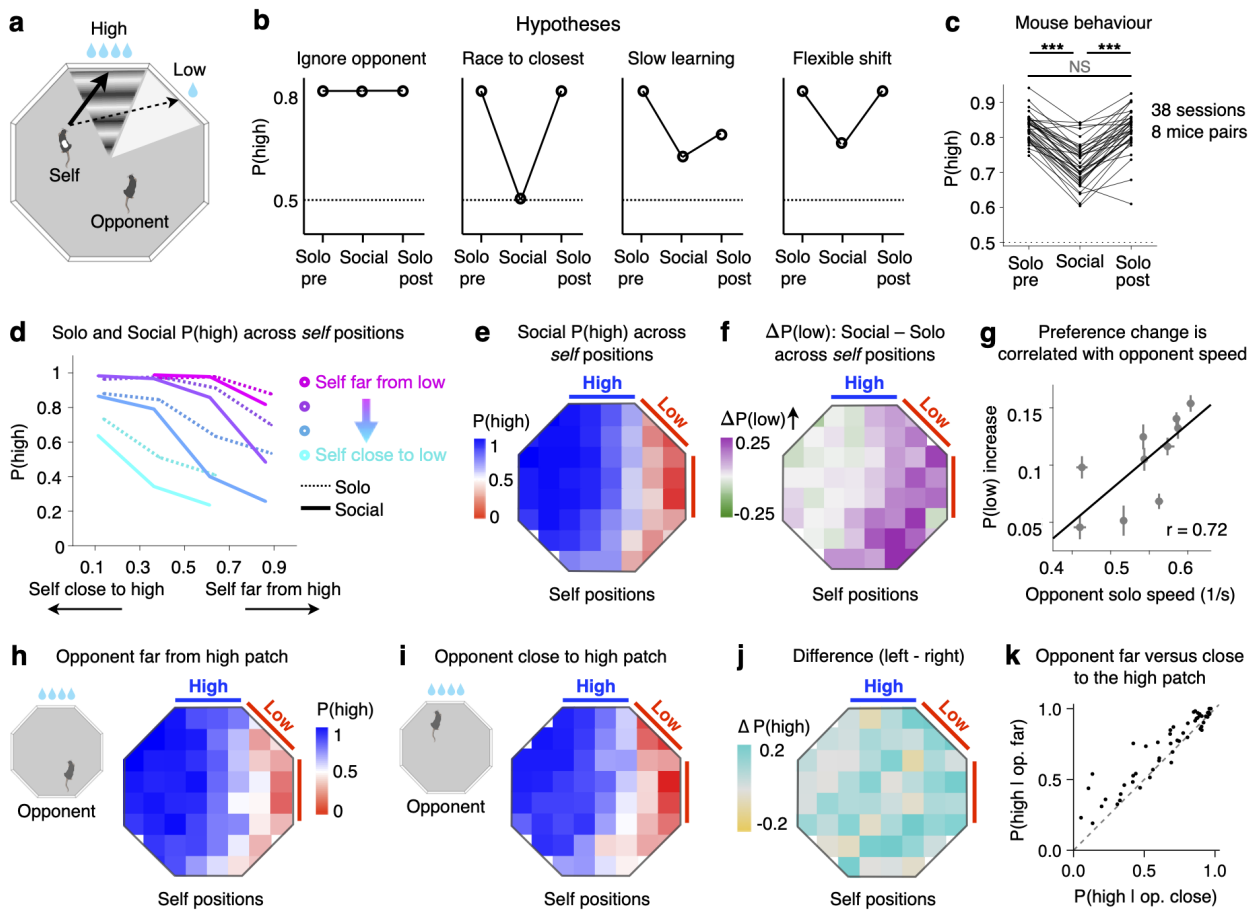

Fig. 2: Mice flexibly shift decision preference under social competition. **a**, Schematic for multi-animal competition in the Octagon task. **b**, Hypotheses for how mice would behave across solo and competitive social contexts. **c**, Session-averaged change in the probability of choosing the high patch across 38 solo-social-solo experiments from 8 mice pairs. Mice decreased preference for the high reward patch under competition. ***, $P < .001$; NS, $P > 0.05$; paired t-tests. **d**, P(high) in solo (dashed) and social (solid) sessions, across varied self distances to the high and low reward patches, similar to Fig. 1**o**. Self is defined as the winning mouse on each trial. **e**, P(high) across varied self start positions in social sessions (n = 38). **f**, Increase in P(low) from the solo context to the social context across varied self start positions. **g**, Correlation between increase in P(low) in social sessions and speed of the opponent in solo sessions. Mice shifted their strategy more when competing against faster opponents. Error bars are s.e.m. across mice and sessions. **h,i**, Similar to **e**, plotted separately for when the opponent was far from the high patch (> 60% distance, **h**) or close to the high patch (< 40% distance, **i**). **j**, Difference between **h** and **i**. Mice chose the high patch more when the opponent was far from the high patch. **k**, Scatter plot based on data from **h** (y axis) and **i** (x axis), each point represents one specific self position bin.

of high reward patch preference in social sessions ($8.58 \pm 0.90\%$ decrease, $P = 1.18 \times 10^{-10}$; Fig. S5c), indicating that speed-accuracy tradeoff was not sufficient to account for the observed changes under social competition.

Why did mice shift their decision preference under social competition? One possibility is that choosing the high reward patch sometimes increases the risk of losing the competition, for example, when 'self' is relatively further from the high patch and when competing against a fast opponent. Under these conditions, it may be reward-maximizing to choose the low reward patch as a safer bet. Indeed, we found that the preference shift in social sessions depended on animals' 'self' distances to the high and low patches (Fig. 2d-f). When 'self' was far away from the high patch, the probability of choosing the safer, low-payout option under social competition increased as much as 24.59% from the solo condition. In contrast, when 'self' was close to the high patch the preference for the high reward option remained the same across contexts (Fig. 2f). We fit a GLMM to predict choices across different contexts and different 'self' distances to goals, and found a significant main effect of the social context across all mice ($\beta_{social} = -0.62, P = 4.45 \times 10^{-27}$; Fig. S6a-f, Table S7), suggesting that social competition shifted animals' preference towards the safer option in a self-position-dependent manner.

Furthermore, we tested whether features of different opponents, such as speed and baseline value preference, modulated strategy shifts. Across different pairs of mice, we found a significant correlation between opponent speed and the magnitude of 'self' preference shift under social competition: the faster the opponent was, the more mice shifted their decision preference towards the low-payout option ($r = .72, P = .02, n = 10$ opponents; $r = .32, P = .0023, n = 88$ sessions; Fig. 2g). On the other hand, opponent P(high) on solo days did not correlate with animals' strategy shift under competition ($r = .04, P = .91, n = 10$ opponents; $r = .01, P = .91, n = 88$ sessions). This result demonstrated that mice shifted their preference flexibly based on specific features of individual opponents.

Finally, we tested if mice could integrate the instantaneous positions of the opponent to guide choices on a trial-by-trial basis. To avoid introducing spurious correlations between self choices and opponent position by only analyzing winner's choices, we also included inferred loser's choices in this analysis (Methods). For visualization, we binned self positions and compared P(high) when the opponent was far (Fig. 2h) or close (Fig. 2i) to the high reward patch at stimulus onset. We found that mice's decision preference was sensitive to opponent positions, and they chose the high patch more when the opponent was far from the high patch (Fig. 2h-k). To quantify this effect, we fit a GLMM to predict self choices as a function of self and opponent distances to goals, and found a significant main effect of the opponent distance to the high patch ($\beta_{p2_dh} = 0.27, P < 10^{-5}$, Table S11) in addition to self distance effects. Moreover, the addition of interactions between self and opponent positions further improved model predictions compared to the main-effects-only model ($P = .003$, Table S12,S13). Thus, opponent position shifted preferences in two ways: first, mice were more likely to choose high when opponents were far from the high patch; second, opponent position modulated animals' sensitivity to self position (Fig. S6g,h).

Interacting artificial agents recapitulate mouse behavior

To gain a quantitative understanding of animals' strategies, we simulated multi-player games with competing reinforcement learning (RL) agents. In particular, we aimed to answer two questions. First, could some of the behaviours observed in competing mice be explained by adapting to the changing reward statistics in social sessions? Second, were animals integrating opponent information to guide choices in a reward-maximizing way? We generated synthetic data from two different types of RL agents, with or without access to opponent position (Fig. 3a,b; Methods). Both types of agents could learn that the relationship between choice and reward changed under competition (since the loser received no reward), but only the 'social' agents, and not the 'socially blind' agents could use opponent position to guide choices. Simulating these distinct artificial agents provided synthetic data to benchmark socially-guided strategies against non-social reward learning in mice behavior.

To approximate a state-action-value function that generalized across the high-dimensional state space in the Octagon task, we used deep Q-networks (DQN) with convolutional layers followed by densely connected layers (Fig. S7a) [40]. On each trial, the input to the DQN was a bird's-eye view of the Octagon arena with positions of the high and low patches as well as self-position (Fig. 3a). 'Social' agents also had access to the positions of the opponent whereas 'socially blind' agent did not (Fig. 3b). The output of the DQN consisted of Q values of 9 possible actions: to choose one of the 8 patches or to abort. Based on the Q values, an action was chosen for that trial, and the corresponding reward was delivered. Reward depended on the value of the chosen patch, discounted by action cost and temporal discounting (Fig. S7b; Methods), and in competitive sessions, depended on winning. The action cost and discount factor were picked to qualitatively match mice solo data (Fig. S7c). The update step of the DQN involved sampling from a 'replay buffer' of past experiences. We used a recency weighting factor, rw, that biased learning from recent experiences. When rw was small, the buffer was sampled more evenly, and as rw got larger, more recent experiences were oversampled (Fig. S8a).

Similar to mice, DQN agents trained in a solo context produced a comparable value preference that generalized across varied start positions and patch configurations; and was modulated by agents' relative distances to the patches (Fig. 3c; n = 20 agents). After solo training, pairs of 'social' or 'socially blind' agents were trained in interleaved solo and competitive multi-agent sessions. For each trial under competition, the positions of reward patches and the two agents were randomly selected, and each agent would choose an action based on that state. The winner was determined by comparing the distance between each agent's start position and its chosen patch to compute which agent would get to its choice first. The winner got a discounted reward, as in the solo context. The loser did not receive a reward. Surprisingly, we found that simple metrics such as the probability of choosing the high reward patch across solo and competitive contexts were not always sufficient to distinguish 'social' versus 'non-social' strategies. When agents learned from a long history of trials ($rw = 0.001$), only the 'social' agents displayed flexible shift of value preference across contexts (Fig. 3d,f).

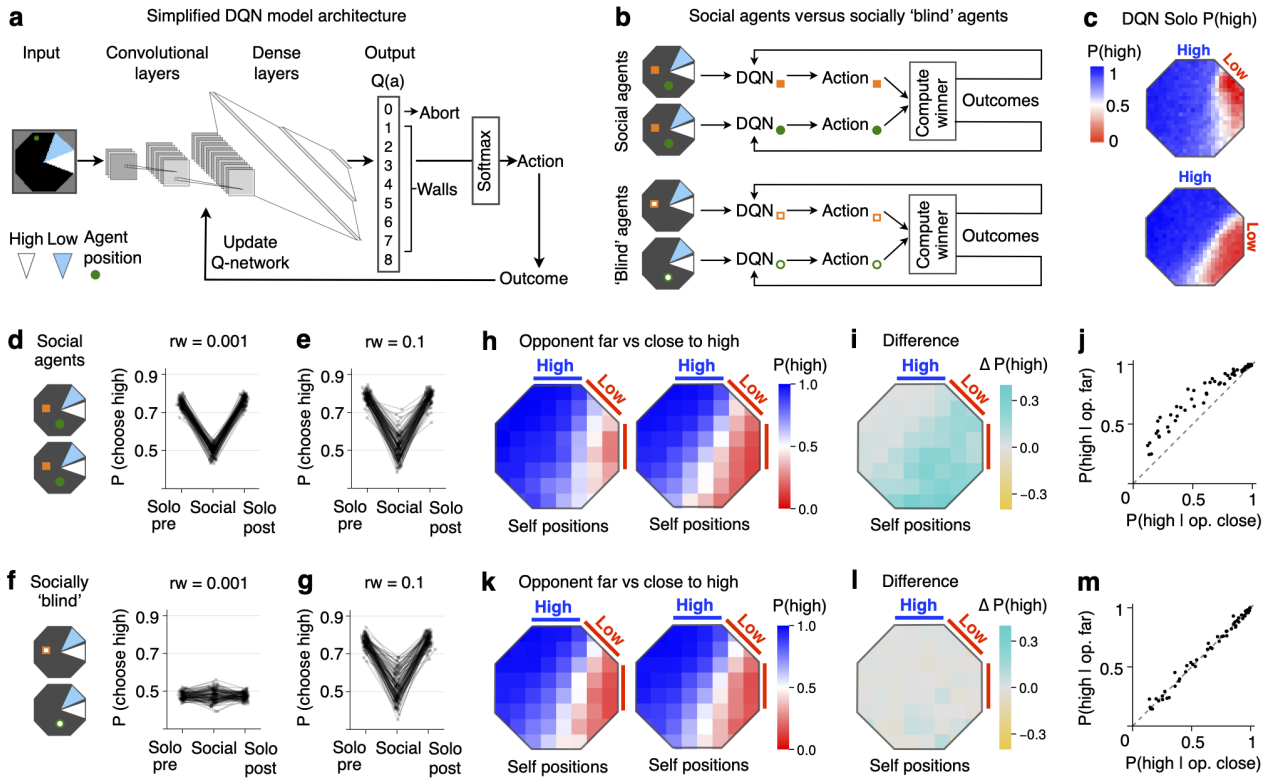

Fig. 3: Simulations of multi-agent competition in interacting DQNs. **a**, Model architecture and schematic trial structure for training and updating. **b**, Two types of DQN agents: 'social' agents have access to position information of self and the opponent whereas 'socially-blind' agents only have access to self information. Only the winner gets rewarded on each trial (Methods). **c**, DQN performance on solo sessions, averaged across 20 agents, rw = 0.1, similar to mice solo performance in Fig. 1**l**,**n**. **d**,**e**, Change in the probability of choosing the high patch across solo and competitive contexts for 'social' agents, with slow (**d**) and rapid learning (**e**). **f**,**g**, Similar to **d**,**e**, for 'socially-blind' agents. With rapid learning, averaged simple metrics could not differentiate the two types of agents. **h-m**, 'Social' agents (**h-j**), but not 'socially-blind' agents (**k-m**), show opponent position sensitivity in choice preferences under competition, even with rapid learning (rw = 0.1).

However, when agents preferentially learned from more recent history ($rw = 0.1$), both types of agents showed 'flexible' shifts across contexts (Fig. 3e,g, Fig. S8), suggesting that rapid non-social learning could generate the type of average preference changes as observed in competing mice. Similarly, self-position-specific preference shifts could also be explained by rapid non-social learning (Fig. S9).

In contrast, opponent sensitivity was uniquely associated with 'socially'-guided strategies, even at the fastest timescale of learning (rw = 0.1; Fig. 3h-m). Although 'socially blind' agents did not have access to opponent information by design, it was important to demonstrate that sensitivity to opponent behaviours could not arise from spurious correlations with non-social factors, (e.g., removing correlations between probability of winning and opponent position by including winner's and loser's choices), providing a valid null hypothesis. The strategies of the 'social' DQN agents, by definition, maximize reward. Thus, quantification of their strategies using the same GLM as we previously used to quantify animals' strategies provided us with a benchmark to assess whether mice were reward maximizing as well. Consistent with mouse behaviours, 'social' agent choices were significantly modulated by the opponent distance to the high patch ($\beta_{p2_dh} = 2.1, P < 10^{-99}$; Fig. 3h-j, Table S15). Furthermore, the effects of opponent positions also depended on specific self positions in the octagon ($P < 10^{-22}$; Fig. 3i, Table S17). As expected, 'socially blind' agents did not show any opponent sensitivity ($\beta_{p2_dh} = -0.03, P = .59$; Fig. 3k-m, Table S19, despite general preference shifts across contexts (Fig. 3g).

Together, these simulations confirmed that the opponent-guided strategies observed in competing mice are reward-maximizing, and cannot simply be explained by non-social learning from environmental changes. However, there are several limitations to DQN as a modelling framework to understand the biological basis of decision-making. First, the DQN are a black box. Dissecting the features and representations that the DQN use to guide choices would require considerable work which is tangential to our efforts here. Second, there is no

noise nor dynamics in the DQN: the output layer is a deterministic function of the input (Fig. 3a) and the DQN does not have reaction times. To overcome these limitations and provide a foundation for interpreting future neural recording and perturbation experiments, we turned to a well-established model of decision-making in neuroscience: the bistable attractor [41–43].

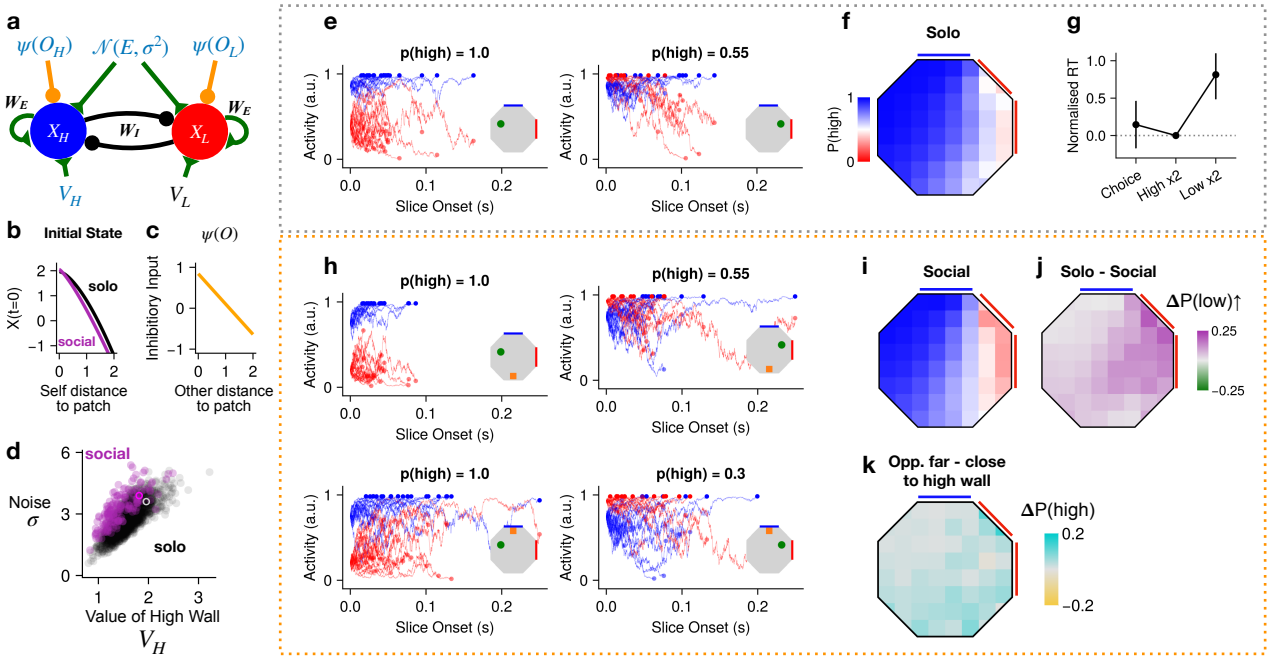

Fig. 4: Dynamical model of choice. **a**, Decision dynamics were simulated by a two-node system (X_H,X_L) representing the decision to choose the high or low patch. V_H, V_L represent the values of the high and low patch. W_E, W_I represent the weight of self-excitation and mutual inhibition. $\psi(O)$ represents the influence of the opponent in the social context. $\mathcal{N}(E, \sigma^2)$ represent Gaussian noise. Blue parameters were fit to data. Triangles and circles represent excitatory and inhibitory connections respectively. **b**, The initial activity level of the two nodes was determined by the agent's distance to each patch. **c**, The distance of 'other' to each patch modulated the respective node. **d**, Comparison of the noise parameter, σ, with the value of the high patch, V_H, for the top 2.5% of fits to the solo (black) and social (purple) mouse data. White and magenta circles indicate the best solo and social fits, which were used to generate the plots in **e-k**. **e-g**, Simulations generated using the best solo parameters. **e**, Trajectories ($n = 20$) of the solo dynamical system for a trial configuration where the agent is close to high (left) and another where the agent is close to low (right). The inset octagon indicates the position of the agent (green circle) and the high (blue) and low (red) walls. The title indicates P(high) for the 20 simulated trajectories. **f**, Solo preference plot for 45° and 90° trials combined (compare with Figs. 1l,n). **g**, Simulated reaction times for the choice trials (with one high and one low patch) and for forced trials (High ×2 and Low ×2; compare with Fig. 1f). **h-k**, Simulations generated using the best social parameters. **h**, Trajectories ($n = 20$) of the social dynamical system. The inset shows the trial configuration as in **e** but with the opponent position at slice onset shown (orange square). **i**, Social preference plot for 45° and 90° trials combined (compare with Fig. 2e). **j**, Social - Solo preference (compare with Fig. 2f). **k**, Influence of opponent position on agent choice (compare with Fig. 2j).

A dynamical model of flexible decisions under multi-agent competition

While watching the behavior of mice in the Octagon during solo sessions, we noticed that when the mice were very close to the low patch, they seemed to be 'drawn' to it. This inspired us to model the mouse behavior as a bistable attractor where the basin of attraction for each patch was modulated by the value of the patch (Fig. 4a). This simplified the decision of the mouse into a binary choice: choose the high-reward patch or the low-reward patch [44]. We used a combination of global optimaization methods and gradient descent to find parameter sets that best explained solo and social experimental data respectively.

The initial state of the dynamical system, on each trial, was based on the animals' distance to the patches at stimulus onset (Fig. 4b). For example, when the animal was close to the high patch (Fig. 4e left), the initial activity of the high node (X_H) was close to the decision threshold, resulting in a strong preference for

high. Conversely, if the mouse was close to the low patch, then the activity of the low node (X_L) was close to threshold, leading to some fast choices for the low patch (Fig. 4e-right). However, in the latter case the noise sometimes allowed the state to 'escape' from the shallower low-patch attractor to the deeper high-patch attractor . Using distance-to-patch as the initial conditions of the dynamical system recapitulated the distance modulation of value preference (compare Fig. 4f with Fig. 1l,n), without explicitly adding a time or action cost as we did in the DQN simulations.

The effect of the opponent on decisions was implemented by a distance-weighted input to the nodes (Fig. 4a,c). For example, when the opponent was far from the patches (Fig. 4h-top), the behavior of the social model was similar to the solo model (compare with Fig. 4e). However, when the opponent was close to the high patch, this had a big influence on dynamics and shifted choices to the low patch, especially when self was close to low (Fig. 4h-bottom). In addition to the opponent input to decision nodes, we found several differences between the best social and solo models: decisions under competition were noisier and less driven by value (Fig. 4d); and the initial state of the dynamical system was slightly more sensitive to position in the social model (Fig. 4b). These mechanisms were sufficient to account for the key experimental findings, including the self-position-specific preference change across contexts (compare Fig. 4j with Fig. 2f) as well as the interaction between self-position and opponent sensitivity under competition (compare Fig. 4k with Fig. 2j and Fig. 3i). Thus, a bistable attractor model that integrated self-position as the initial condition, with value and opponent inputs, captured the behavioral phenomena observed in the mice.

These models make several predictions. On the behavioural level, even though the model was only fit to the choices between the high and low reward patches, we could simulate forced trials where both nodes received the same value input and estimate the corresponding reaction times (RT). We found that RTs in the model, like the mice, were fastest when choosing between two high patches, slowest when choosing between two low patches, and intermediate when choosing between high and low (compare Fig. 4g with Fig. 1f). The slow RT when choosing between two low patches may reflect a process where noise pushed the system back and forth between two shallow basins of attraction. On the neural level, a biological implementation of this dynamical system would have several features. First, we would expect to observe neurons that are selective for abstract value-guided choices invariant to the specific patch locations on each trial. Second, self distances to patches should be easiest to decode at the time of stimulus onset and decay as the dynamics settle into one of the fixed points. Third, opponent distances to patches should also be decodable from the choice neurons. Finally, The dynamics of choice coding can be predicted for each trial configuration: comparing the left panels of Fig. 4e,h we see that the model predict 100% choices to the high patch, but the degree of overlap between red and blue traces varies considerably. This makes an additional prediction regarding the robustness of the dynamical system to perturbations.

Discussion

In this study, we established a novel, freely-moving mouse paradigm to investigate value-based decision-making in a fast-changing, competitive social environment. The dynamic nature of our task encouraged mice to develop flexible and generalizable value-guided strategies across varied trial configurations (Fig. 1). This allowed us to disentangle the effects of value and space (distances to goals) on solo decision preference (Fig. 1k-o), and reveal how animals traded off these factors under competition (Fig. 2c-f). Future studies with more detailed characterization of behavior, such as field-of-view and trajectory [45–47], may reveal further embodied modulation of individual choices. Beyond decision preferences, we found that mice formed higher-order strategies, such as returning to the center of the arena before each trial, potentially to minimize travel time (Fig. S4). Strikingly, control mice that learned to associate equal value with both visual patches developed different start position patterns that matched the respective optimal solutions, in addition to distinct decision preferences (Fig. S3,S4). The discovery of this higher-order strategy underscores the importance of studying decision-making in settings where animals' intrinsic behavioral variance is unrestricted.

Our social foraging game is different from previous rodent competition tasks in several ways [48–50]. First, in the Octagon, there were two options to choose from on each trial, and the reward-maximizing strategy during social competition involved choosing the low reward option when appropriate. This allowed us to characterize how animals shifted their decision preference to maximize reward under competition (Fig. 2b,c). Second, the locations of the reward patches were unpredictable on each trial. As a result, it was not optimal to use a spatial strategy of occupying a specific part of the arena to monopolize the high reward source. Although we did not explicitly measure the effects of social rank on competitive foraging, we predict that dominant animals may have less physical advantage in our game compared to previous tasks where reward locations are stable within blocks or sessions [48, 49]. We observed that mice shifted strategies more when competing against fast opponents (Fig. 2g), reminiscent of past reports where submissive mice adjusted their behaviours more than dominant mice in a simple rank test [51]. Future studies that directly compare social foraging in stable versus dynamic environments may reveal context-dependent strategies and differential involvement of frontal cortex under competition. Finally, because mice actively surveyed their environment before each trial (Fig. 1i), animals' relative distances to both patches and in turn the corresponding reward probabilities varied across trials. This notably increased the complexity of the decision process. To our surprise, mice integrated the instantaneous positions of both 'self' and 'other' to make rapid yet highly flexible choices on each trial (Fig. 2f-k).

Using multi-agent RL simulations [52], we confirmed that the opponent sensitivity observed in mice is consistent with maximising discounted rewards and could not be explained by fast or slow non-social learning (Fig. 3h-m). With noiseless information about each trial, the 'social' agents were more strongly influenced by opponent position than mice. Future work could examine whether RL agents, with inputs that more closely reflect the sensory experience of mice, would exhibit behavior that quantitatively matched mouse behavior. On the other hand, we found that some changes between solo and social sessions could be explained by learning alone (Fig. 3d-g, Fig. S9). This stresses the importance of using artificial agents to generate synthetic datasets, which can be used to identify benchmarks for hypothesis testing; especially in tasks like the Octagon, where key variables are outside of experimental control. For example, initial analysis of 'socially-blind' DQN behavior seemed to indicate that they were sensitive to opponent position: an impossibility. We realized that by only including the winner's choices in the analysis, we introduced a spurious correlation (Table S23): because the probability of winning depended on the position of opponent. This finding from synthetic data analysis drove us to use video tracking to infer the loser's choices, so that we could avoid this pitfall.

The dynamical model simulations gave us insight into the decision process on each trial and how this process was influenced by solo versus social context and by trial configuration (i.e., the distance of each mouse to each patch). While bistable attractors have been used extensively in neuroscience as a framework to understand decision-making [41–43, 53, 54], and dynamical systems have been used to describe social behavior [44, 55–58], we are not aware of any previous work that has used them to model behavioural data during social competition for resources. In contrast to the RL simulations, where we discounted rewards by time and distance, we found that attractor dynamics was sufficient to account for animals' tendency to choose the low wall when they were close to it. The relationship between delay-discounting and attractor dynamics is relatively under-explored [but see 59]. It would be interesting to quantify what characteristics of bistable attractors could give rise to changes of mind, e.g. 'dynamic inconsistency' in intertemporal choice, that are normally explained using hyperbolic discounting functions [60].

Fitting the dynamical system to data revealed that in addition to the effect of the opponent, decisions under competitive pressure were noisier and less driven by option value, in turn making decisions more sensitive to initial positions. The dynamical system framework can be used to generate synthetic 'neural' data and make specific predictions about electrophysiological finding, such as the relative decodability of different task variables across time in a trial. Moreover, the distinct inputs to the model (value information, opponent information) are concrete neural candidates for investigation. For example, the nucleus accumbens and amygdala have been suggested to provide value information to the prefrontal cortex in social decisions [61]. Our behavior and dynamical framework would be ideal for quantitatively testing the relative contributions of these inputs. The current implementation of the model interprets the dynamics as a process that reaches a decision threshold followed by a ballistic movement to the goal. A future extension could better integrate the embodied nature of the task by equating the state of the dynamical system with animal's continuous trajectory as they move to the patch [62].

The neuroscience community is increasingly interested [63, 64] and equipped [65–69] to understand innate and learned social behaviours [70–74]. Here, we extended behavioural, analytical and computational methods classically used in solo decision-making studies to multi-agent competition, providing a novel framework towards mechanistic understanding of decisions in a fast-changing and social environment.

Methods

Subjects.

All experiments were performed under the UK Animals (Scientific Procedures) Act of 1986 (PPL PD867676F) after local ethical approval by the Sainsbury Wellcome Centre Animal Welfare Ethical Review Body. Ten adult C57BL/6J mice (8 males and 2 females; Charles River Laboratories), at least 8 weeks old at the start of the experiments, were used in this study. Mice were co-housed with their litter mates, with ambient temperature and humidity, and 12-h reversed day-night cycles. Mice were water-restricted before the start of behavioural training, with *ad libitum* access to food. On training days, mice received water as reward inside training rigs. Supplementary water was provided for mice to maintain a stable body weight (around 90%, but at least 85% of their starting body weight).

Behavioural apparatus.

Hardware. All experiments were conducted in custom-designed and manufactured open arenas (Fig. S1; NeuroGEARS Ltd., UK). Each arena consisted of an elevated base and 8 white acrylic wall panels, forming an octagonal space 80 cm in diameter and 50 cm in height. To allow visual stimuli presentation on the arena floor [75], a short throw projector (HD29HST, Optoma) was used to project images onto a tilted mirror on the room floor, which then reflected the visual stimuli onto the arena base. The arena base was made from a transparent acrylic sheet with a rear projection film (Pro Diffusion, Pro Display) glued underneath, located 70 cm above the floor. Calibration was conducted using custom software in Bonsai [66] to ensure approximate equal luminance across the arena base and sufficient dynamic range for visual stimuli presentation. Each arena wall panel contained two nose-poke ports (OEPS-7246, Open Ephys), 3 cm above the arena base, for nose-poke detection upon infrared (IR) beam breaking. Each port also had an LED for additional visual aid; as well as a water spout, which was connected to a water reservoir via a computer-controlled pinch valve (161P011, NResearch) for reward delivery. To facilitate real-time mouse position tracking in the dark, custom made infrared (IR) illumination strips (850 nm LED) were mounted above, around, and below the arena. Videos were captured by an IR camera (BFS-U3-16S2M-CS, Flir; LN048 lenses) at 50 fps, mounted 124 cm above the centre of the arena base. A second monochrome camera with a bandpass filter (2020OFS-585, OFS) was used to monitor visual stimuli and animals' ongoing behaviours.

Data acquisition and task control. Custom written software in Bonsai [66] was used to read and control all sensors and effectors in the arena via Harp devices (OEPS-7240, Open Ephys; Fig. S1a). In brief, each arena wall's nose-poke ports and water valves were controlled by one Harp device (8 devices in total); and all 8 devices were synchronised using 2 clock Harp devices (OEPS-7231, Open Ephys). Projector visual stimuli presentation and camera video acquisition also triggered and synchronized by Bonsai. Behavioural task logic was written in Python-based software; and an Application Programming Interface (API) was developed to communicate between Python and Bonsai using a messaging library (https://github.com/glopesdev/pyOSC3).

Data acquisition, video synchronization, data storage and access format were based on an API developed by the Aeon Foraging Behaviour Working Group [76].

Behaviour.

Solo task. In the final stage of the solo task, mice (n = 8) were guided by visual stimuli to perform value-based decision-making in the Octagon arena. The start of each trial was cued by a background colour change in the arena floor from black to gray. After a variable delay (0.5–1.5 s), two floor-projected visual stimuli (grating or white light triangular slices) were presented to indicate the location and value of the active reward patches for that trial. One visual pattern was associated with a high reward (3–5 units of 2.5 µL water reward, $p = 1$), and the other pattern was associated with a low reward (1 unit reward, $p = 0.3$–0.6). The luminance of the two visual stimuli were balanced to avoid perceptual biases; and the mapping between visual stimuli to reward sizes were counter-balanced across mice. Mice always competed The stimuli remained on until mice indicated their choice by nose-poking into one of the available patches within a response window (30 s from stimulus onset), which in turn triggered the corresponding reward. Poking into either of the two nose-poke ports on a single wall resulted in identical outcomes. Failure to respond within the 30 s response window was considered a 'miss' trial. Response time (RT) was defined as the time between stimulus onset and animals' choice; and only trials with RTs shorter than 15 s were included in further analyses (mean RT: 1.77 s). All trials ended with an intertrial interval (ITI) of 6–12 *s*. Mice were not constrained to be at a particular location in the arena before each trial start. To increase the probability of mice attending to visual stimuli on the next trial, a new trial would only start after a minimum of 2 s where mice were not poking into any of the 8 walls. In each session, 80% of trials were 'choice trials' where mice chose between two different reward options; and 20% were forced trials where mice chose between two identical reward options in different locations. The locations of patches were randomized across trials, and the two patches could be 45° apart (neighbouring), 90° apart, or 135° apart. Sessions ended when animals showed increased miss rate as a sign for satiation. A control group of mice (n = 2) were trained to associate the grating and light floor stimuli with equal reward sizes (3 units); and their solo preferences for different stimuli were characterized and compared to the experimental group.

Training procedure. In the first behavioural session, mice were habituated with the arena and learned to poke into LED-illuminated ports for water reward. To aid operant conditioning in this large arena, all 16 ports were illuminated on each trial to start with and poking into any of the ports within a 60 s response window would trigger a reward (3 units reward). The number of available ports on each trial gradually reduced upon every 10 successful reward collection, until only two ports from 1 randomly chosen wall were illuminated on each trial. Mice typically learned to follow LED-illuminated ports in the Octagon arena within the first session. In the next 1–2 sessions, mice were trained to follow floor-projected visual stimuli rather than port LEDs for reward. On each trial, a single triangular stimulus (grating or white light slice) would appear on the arena floor to cue a specific wall as the active reward patch for that trial. After a variable delay (7–15 s), if no response in the active patch was registered, the corresponding port LEDs would light up to facilitate transfer learning. When mice could reliably respond to floor stimuli without port LEDs, only floor stimuli would be presented on each trial, and a response within a 30 s window would trigger the corresponding reward for that visual stimulus. To reduce anxiety, cage mates were trained together in these initial shaping steps, after which mice started solo training to discriminate visual stimuli associated with different reward sizes. As described above, two floor stimuli were presented on each trial and mice needed to choose one of the two active patches for reward. Poking into any non-active patch (extremely rare) was recorded but not punished or rewarded. Animals with stable preference for the high reward patch (> 70–80%) then moved on to competitive social sessions.

Competitive social task. The trial structure in the competitive task was identical to the solo task. Two well-trained mice (cage mates) were placed in the arena together. On each trial, only the first mouse to report its choice would be rewarded for their choice; and the slower mouse would receive no reward. Similar to the solo task, new trials would only start after a minimum of 2 s where neither of the two mice was poking into any of the 8 walls. On each trial, the winner mouse that received the reward was referred to as 'self' and the slower mouse was referred to as 'other' or the opponent; and 'self' identification varied across trials. Only cage mates with the same sensory-reward mappings were paired together for competition. For each pair of competing mice (n = 8 pairs), we conducted 3–8 triplets of solo-social-solo experiments. Social sessions lasted longer than solo sessions to approximately match the number of rewarded trials per mouse (social 118 versus solo 151 trials/mouse/session).

Video tracking and identification.

To track mouse identification during multi-animal sessions, we bleached back fur in one of the two competing mice so they could be easily distinguishable in real time by Bonsai tracking algorithms and post-hoc by SLEAP [65]. Briefly, mice were anesthetized with isoflurane (2%). Human hair bleach was mixed according to the manufacturer's instructions; and applied only to the top of animals' fur to avoid skin irritation. The bleach stayed on for 7 min before it was wiped off and rinsed with saline. Mice were then placed back to a heated cage to recover from anesthesia. Fur bleaching was always conducted before solo sessions to ensure that there was no change in baseline task performance after brief anesthesia or the bleaching procedure. Bleaching was repeated when needed, usually after 2-4 weeks.

The infrared (IR) camera was used to track animals' positions and identifications in real time at 50 fps. Briefly, the IR camera imaging frames were thresholded to expose the region of interest (ROI) corresponding to the mouse (the black area of the mouse's fur versus the white arena floor). In social competition experiments, the bleached fur reduced the ROI area of the bleached mouse, allowing the two mice to be distinguished based on the ROI size (Fig. S1d). On each frame, the central position (centroid) and area size of a large ROI (non-bleached mouse) and a small ROI (bleached mouse) were tracked and recorded. We used the centroid positions and area sizes at the time of response to determine which mouse was the winner ('self') on each trial. Tracked body size could vary depending on animals' varied postures within a session. To ensure correct identification using our method, we plotted the distributions of tracked body sizes across all the frames in each pair of solo control sessions. If the two distributions were non-overlapping, we proceeded with the social experiment on the next day; otherwise bleaching was repeated. Identification tracking was further validated post-hoc using SLEAP [65].

After extracting animals' positions from the custom Bonsai pipeline, the trajectories underwent further offline processing, to fill any gaps in the positional data using linear interpolation, and to smooth outlier data due to tracking artifacts. Although the loser's choice was not rewarded, a subset of analyses required us to include both animals' choices to eliminate spurious correlations. We analyzed post-processed trajectory data and inferred the loser's choice by estimating which of the two active walls they were closer to, 1 second after the winner's choice. We only included trials where the loser's distance was less than 12 cm from the choice wall, so trials where they were at a different response port (e.g. still at a reward port from a previous trial) were excluded. Using a stricter criterion did not change any of our main conclusions.

Behavioural data analysis.

Solo behaviours. Analyses of behavioural data were conducted using custom software written in MATLAB (Mathworks, 2021b) and Julia [1.10; 77]. For all analyses, we used a coordinate system where the center of the octagon was [0,0] and the distance to the midpoint of the patch walls was 1. To characterize post-learning performance (Fig. 1), we included the last 30 days of stable solo training from each mouse. Generalization of value preference across patch locations (Fig. 1h) was demonstrated by concatenating all the choices from 8 mice, and plotting the probability of choosing the high patch, P(high), as a function of actual high wall locations. To quantify response time (RT) differences across trial types (Fig. 1f), we calculated the median RT for each trial type per mouse, and plotted the log ratio of RTs, normalized to trials where mice were 'forced' to choose between two high patches. Within-subject bootstrapped tests were used for significance testing. The relationship between forced trial log RT ratio and choice trial value discrimination was quantified using linear correlation across sessions (Fig. 1g). To test whether the high/low mice were statistically closer to the arena center at trial start, we used the Receiver operating characteristic (ROC) analysis to distinguish trials from the high/low mice versus the control mice. We estimated the area under the ROC curve (AUC) between concatenated trials across mice groups, or between each high/low mouse against the control group.

We quantified the spatial effects on animals' value preference in two ways. First, for visualization, we rotated and flipped varied trial configurations so that the high reward patch was at direct north, and the low reward patch was east to the high patch (Fig. 1i-k,m). Trials were then binned based on start positions relative to the high and low patches, and averaged P(high) was plotted for each start position bin in the Octagon space (Fig. 1l,n). Second, for statistics and model comparisons, we calculated the distance from the mouse's start position to the high patch and to the low patch on each trial, and then used distances rather than positions to evaluate spatial effects on performance (Fig. 1o). The first approach was visually salient, whereas the second approach allowed us to combine trial conditions where the two patches were 45° or 90° apart.

To assess animals' overall solo strategies, we fit generalized linear mixed models (GLMM, implemented in MATLAB function `fitglme`) where animals' choice on each trial was a logistic function of intercept and their distances to the high and low patches at stimulus onset as fixed effects, as well as within-subject random effects (See Statistical Appendix for details). We compared 4 models: model 3 (Table S3) includes distance to low only; model 4 (Table S4) includes distance to high only; model 2 (Table S2) includes both distances as regressors; and model 1 (Table S1) includes both distances and their interaction (Fig. S2c). Twenty-fold cross-validation was used to calculate per trial normalized log likelihood for model comparisons (Fig. S2e), where P values were calculated using paired bootstrapped tests across folds. Significant effects of distances and their interaction on P(high) were quantified based on regressor coefficients and nested model comparisons. We also fit model 2 to each animal separately and reported the regression coefficients per animal (Fig. S2d). All regression tables are shared in the Statistical Appendix.

Competitive behaviours. To assess how competition shifted animals' overall decision preference, we calculated P(high) across the winning choices in each social session, and compared that to P(high) on the corresponding solo sessions one day before and one day after (Fig. 2c). For each pair of animals, trials from both animals' solo sessions were concatenated to calculate the average solo P(high) for that day. Paired t-tests were used for statistical testing across 38 session triplets. Further separating trials from each social session into the first and second halves did not reveal any consistent change due to slow adaptations within the social sessions (Fig. S5d). To ensure that the preference shift observed was not merely a speed-accuracy tradeoff, we conducted a control analysis that simulated selection of the faster choice between each pair of mice without the element of competition. Specifically, for each trial in the social session, a pair of trials were randomly chosen, with replacement, from two animals' solo control sessions 1 day before; and only the faster trial from each pair was included into an RT-corrected solo control session (Fig. S5a-b). We found similar results using this conservative control analysis.

To visualize how animals' value preference shifted under competition as a function of self positions relative to goals, we used similar methods as described for Fig. 1k-o, but considering whether the trials came from solo vs. social sessions (Fig. 2d-f). For statistics, we fit GLMMs on concatenated solo and social trials, where animals' choice on each trial was a logistic function of the chooser's distances to the high and low patches at stimulus onset as well as whether the trial was from a solo or social session (social context) as fixed effects, and within-subject-pair random effects (See Statistical Appendix for details). We compared 3 models: model 3 (Table S8) only included self distances to goals and their interaction as regressors (Fig. S6d); model 2 (Table S7) included self distances to goals and the social context (Fig. S6c); and model 1 (Table S6) included additional interactions between self distances and the social context (Fig. S6b). Twenty-fold cross-validation was used to calculate per trial normalized log likelihood for model comparisons (Fig. S6f), where P values were calculated using paired boostrapped tests across folds (Table S9). Significant effects were quantified based on regressor coefficients and nested model comparisons. We also fit model 2 to each animal separately and reported the regression coefficients per animal (Fig. S6e).

To test whether mice shifted their strategies differentially when playing against varied opponents, we extracted choices by each mouse in the social session, and compared how P(high) changed from their respective solo performance the day before as a function of opponent features. Specifically, we estimated opponent speed (1/response time) and opponent preference for the high reward patch on solo days, and correlated these opponent features with how much 'self' preference was shifted when competing against these opponents using linear correlation. We found that animals shifted their strategy more against fast opponents (Fig. 2g). To ensure this correlation did not result from analyzing winners' choices only, we included inferred losers' choices and still found a significant correlation between opponent speed and 'self' strategy shift in social sessions ($r = .24, P = .04, n = 70$ sessions).

Finally, we characterized how opponents' instantaneous positions relative to goals modulated animals' choices on a trial-by-trial basis (Fig. 2h-k). To avoid any spurious correlation between the probability of 'self' winning and opponent's positions across trials (Table S23), it was essential that we included both the winners' choices and the inferred losers' choices for this analysis. For visualization, we concatenated all social trials and separately plotted two conditions across all 'self' positions: when the opponent was far from the high patch ($>$ 60 percentiles of all opponent distances; Fig. 2h) and when the opponent was close to the high patch ($<$ 40 percentiles; Fig. 2i). We also plotted the mean difference for each self position bin (Fig. 2k). For statistics, we used GLMMs (using MixedModels.jl [78]) to model each social trial as a logistic function of the chooser's distances to the high and low patches at stimulus onset as well as the opponent's distance to the high patch as fixed effects, and within-subject random effects (See Statistical Appendix for details). We compared

3 models: model 3 (Table S12) only includes self distances to goals and their interaction as regressors; model 2 (Table S11) includes self distances to goals and the main effect of the opponent distance to the high patch; and model 1 (Table S10) includes additional interactions between self distances and the opponent distance to the high patch. Significant effects were quantified based on regressor coefficients and nested model comparisons (likelihood ratio test). The best fit model (model 1) was visualized in Fig. S6g,h.

Simulations of optimal start positions.

To determine optimal start positions (Fig. S4), we assumed that agents would want to minimize the time (and therefore distance) from the start position to the desired patch. We performed these simulations in Julia (1.10) using the Meshes.jl package to define an octagon with an inradius (distance from the center to the midpoint of each edge) of 1. Then, we created a grid of start position points (200×200) uniformly spaced from [-1,1] and determined the points that were inside the octagon. Of the 40000 points, 32756 were inside the octagon. We found that this number of start position simulations generated a smooth visualization [79].

For high/low mice to maximize water intake in solo sessions, animals almost always chose the high reward patch. In order to minimize the distance to the high patch across trials, the best start position would be the one that had the shortest average distance to a single randomly appearing wall, so we computed the mean distance from every point in the octagon to a randomly appearing wall (Fig. S4a). If a mouse instead wanted to choose the closer one of the two available walls, as in 'equal' mice, then the optimal start position is the position in the octagon with the shortest average distance to either of the two randomly appearing walls (for both 45° and 90° trials; Fig. S4b,c).

Reinforcement Learning: DQN model.

The reinforcement learning (RL) agent employed Q-learning, a form of off-policy temporal difference control algorithm that learned to estimate the action value function through iterative updates. In general, the goal of RL is to learn a policy, π, that maximizes the expected cumulative reward over time. Following the approach of Mnih et al., [40] we implemented a Deep Q Network (DQN) to approximate the optimal action-value function:

$$Q^*(s_t, a_t) = \max_{\pi} \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i r_{t+i} \mid s_t, a_t, \pi\right] \tag{1}$$

where Q^* is the maximum sum of all the future rewards r_t, discounted by γ at each time step t, if the agent follows the policy π.

Our environment was a 64×64 pixel image (Fig. 3), providing a bird's-eye view of the Octagon arena. The agents were depicted as simple geometric shapes: circles and squares. The action space consisted of the 8 possible wall choices in the arena, along with a null action (i.e., an action which avoids action costs but gets no reward). Rewards, R for each agent depended on their position within the environment. Each action was associated with a high, low, or no reward based on the configuration of the walls in a given episode (i.e., trial configuration). Additionally, the reward was discounted by γ^d and an action cost, AC, was subtracted from the reward, representing the effort required for the agent to traverse the Octagon arena and reach the selected wall. This action cost was calculated under the assumption that the agent followed a direct path from its starting position to its chosen destination, which is why the cost term is written as the sum of a geometric series. The reward for a specific choice was expressed by the following equation:

$$R = \gamma^d r_{choice} - AC \cdot \frac{1-\gamma^d}{1-\gamma} \tag{2}$$

$$r_{choice} = \begin{cases} r_{high} = 750 & \text{if choice} = \text{high wall} \wedge (\text{solo} \vee \text{won}) \\ r_{low} = 500 & \text{if choice} = \text{low wall} \wedge (\text{solo} \vee \text{won}) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where d was the distance of the agent from it's starting position to the wall chosen (measured in pixels), $\gamma = .98, AC = 5$. In the social context, the winner was determined purely based on the distance of the agents to their choice. The parameters values$(r_{low}, r_{high}, \gamma, AC)$ were chosen to qualitatively fit the mouse behavior in

the solo context, other hyperparameters were described in Table 1. Each episode consisted of a single transition (i.e., a single action to choose the wall), and decisions were modeled solely on the initial positions of the agents at the beginning of the trial. An 'episode' for the DQN corresponded to a single trial experienced by mice in the experimental arena.

The DQN took the image input of the environment and had 2 convolutional layers followed by 3 fully connected layers. The final layer of the DQN represented the Q-value for every available action in that state for the agent, which was used to choose an action by applying a softmax with temperature (τ). We utilized experience replay with a replay buffer of 1,000 episodes. Episodes were sampled using a geometric distribution $G(p = rw)$ to prioritize recent observations, where rw is the recency weight.

Simulation Details In order to facilitate comparison between mice and DQN, we first pre-trained the CNN layers of the DQN to label trials as solo or social trials (details see below). We reasoned that mice know whether or not another mouse is in the arena, so we didn't want this aspect of learning to slow down task training. Likewise, we first trained the DQN in the solo context until their behavior was qualitatively similar to expert mice (5000 episodes, which is similar to the number of trials in 30 sessions for a mouse). Then we alternated 500 episode blocks of solo and social contexts. Since the DQN needed to distinguish 'self' and 'other' from the bird's eye-view input, 10 DQN learned that 'self' was a circle and 10 DQN learned that 'self' was a square. To mimic the individual differences in mice, we trained DQN using 10 random seeds to initialize [80] the dense layers for each agent type (circle and square), yielding 100 pairwise simulations (10 circle vs. 10 square agents). These simulations were repeated (with the same seeds) for social vs. socially blind agents and for three levels of recency weighting (rw; Fig. S8), giving $100 \times 2 \times 3 = 600$ simulations. On each trial, the agents were randomly placed within the arena, ensuring that the glyphs representing the agents in the environment were far apart to avoid any occlusion. The position of the low and high walls was also randomly chosen throughout each episode, maintaining the relative positions between them as in the experiments(45 deg and 90 deg separations) giving 32 possible configurations of the high/low walls. All the simulations and the DQN training was done in Python using PyTorch.

Pre-training Before task-specific learning, the CNN was trained on a binary classification task designed to differentiate between solo and social contexts based on the input image. In the solo context, only one agent (either a circle or a square) was present, whereas in the social context, both agents were included in the image presented to the network. This pre-training enabled the network to differentiate between solo and social contexts without acquiring strategies specific to the decision-making task. Only the CNN layers from the pre-training were used as initialization in the DQN, allowing the model to continue learning and adapt these parameters during the task-learning phase.

DQN Analysis To plot P(high) for the DQN as a function of position in the solo context (Fig. 3c), we took the last 2000 trials from the first solo block for the 20 DQN agents and averaged across agents. For the plots of DQN behavior in the social context (Fig. 3d-m), we took data only from the last social session (n = 500 trials/simulation $\times$ 100 simulations = 50 000 trials). For comparisons across the solo and social contexts in Fig. 3d-g, only the solo sessions before and after the last social session were included.

Hyperparameter	**Value**
Learning rate (α)	0.0001
Softmax temperature (τ)	5
Batch size	128
Replay buffer size	1000
Recency Weight (rw)	0.1, 0.01, 0.001
Loss function	SmoothL1Loss
Optimizer	AdamW

Table 1: DQN Hyperparameters

Attractor modelling.

We simulated a 2-node dynamical system to understand the contributions of a small number of interpretable factors to the decision processes in the Octagon task and to quantify how those factors changed between the Solo and Social contexts. There was no learning in these simulations, so we did not simulate pairs of dynamical systems to determine a winner on each trial. Instead, we simulated each trial of social sessions twice:

once from the perspective of the winner and once from the perspective of the loser (for the trials where we inferred the choice of the loser, as in Fig. 2j). We ran the dynamical simulations in Julia [1.10; 77] using the DifferentialEquations.jl package [81] and visualized the simulations using Makie.jl [79]. The starting point for the implementation was the two-node system fit to data from an accumulation of perceptual evidence decision making task [53]. Here, the activity of the two nodes, X, were governed by the following equations, where V_i was the value of each patch, ψ governed the influence of opponent position, W_E was the strength of self-excitation, W_I was the strength of mutual inhibition, E and σ were the mean and standard deviation of the external drive, and z_i was distance of the opponent to each patch (during social sessions). The initial value of each node, $X_i(t=0)$, depended on the distance of the mouse to the respective patch, d_i:

$$dX_1 = dt(-X_1 + W_E f(X_1) - W_I f(X_2) + V_1 + \psi(z_1) + \mathcal{N}(E, \sigma^2))/\tau \tag{4}$$

$$dX_2 = dt(-X_2 + W_E f(X_2) - W_I f(X_1) + V_2 + \psi(z_2) + \mathcal{N}(E, \sigma^2))/\tau \tag{5}$$

$$f(x) = 0.5 \tanh(x) + 0.5 \tag{6}$$

$$X_i(t=0) = m{d_i}^r + b \tag{7}$$

$$\psi(z) = \begin{cases} 0 & \text{if solo} \\ nz^s + c & \text{if social} \end{cases} \tag{8}$$

The solo model had 6 free parameters and the social model had an additional 3 free parameters (Table 2). Other parameter values were picked based on previous work [Table 3; 53]. We combined data across mice in order to find parameters that captured the overall patterns of behavior. Initial attempts to find the maximum likelihood estimates of the parameters using standard minimization algorithms (Nelder-Mead) revealed that our system had many local minima. To overcome this, we randomly selected parameter sets (10 000 sets for solo and social) from a range and simulated each trial in the data 200 times to get a consistent estimate of the likelihood of the data given that parameter set. We took the best solo and social model from this 'shotgun' approach and then used minimization to find a nearby best model. The best models had a likelihood per trial of 0.6759 for solo trials (n = 24881) and 0.6388 (n = 10460) for social trials.

Parameter	**Best Solo**	**Best Social**
E	1.239	1.916
σ	3.775	3.899
V_H	2.199	1.815
m	-1.155	-1.623
r	1.540	1.238
b	1.471	0.777
n	n.a.	-0.741
s	n.a.	1.000
c	n.a.	1.534

Table 2: Best model parameters.

Parameter	**Value**
W_E	1.5
W_I	4
dt	0.001 s
τ	0.1 s
V_L	0.5

Table 3: Fixed parameters for the dynamical models.

Author contributions

C.A.D. and J.C.E. conceived the project, the arena and designed the behavioral task. J.C.E., B.C., and A.A. contributed to the software and hardware design and construction of the Octagon arena. C.A.D. and I.O.

collected the behavioural data. C.A.D. analyzed the behavioural data, with J.C.E. and M.R's inputs. Q.Y., M.R., and J.C.E. developed and analyzed the DQN simulations. J.C.E. performed the dynamical modelling. C.A.D. and J.C.E. wrote the manuscript, with contributions from M.R. and I.O.

Acknowledgments

This work was funded by the Sainsbury Wellcome Centre Core Grant from the Gatsby Charitable Foundation (GAT3755) and Wellcome (219627/Z/19/Z); by a UK Research and Innovation grant (EP/Y008804/1); and by the Gatsby Initiative in Brain Development and Psychiatry (GAT3955). We thank members of the Duan laboratory and members of the Erlich laboratory for insightful discussions and advice; T. Branco, T. Margrie, and T. Mrsic-Flogel for comments on the manuscript; SWC Neurobiological Research Facility and S. Townsend from FabLabs for technical support.

Competing interests

The authors declare no competing interests.

Code and Data Sharing

Code and data will be shared at the time of publication.

Supplemental Videos

1. Supplemental Video 1. Solo Example Trials. https://shorturl.at/0nUYX
2. Supplemental Video 2. Social Example Trials. https://shorturl.at/R5C01

References

[1] Glimcher, P. W. The neurobiology of visual-saccadic decision making. *Annual Review of Neuroscience* **26**, 133–179 (2003).

[2] Svoboda, K. & Li, N. Neural mechanisms of movement planning: Motor cortex and beyond. *Current Opinion in Neurobiology* **49**, 33–41 (2018).

[3] Gold, J. I. & Shadlen, M. N. The Neural Basis of Decision Making. *Annual Review of Neuroscience* **30**, 535–574 (2007).

[4] Hanks, T. D. & Summerfield, C. Perceptual Decision Making in Rodents, Monkeys, and Humans. *Neuron* **93**, 15–31 (2017).

[5] Padoa-Schioppa, C. Neurobiology of Economic Choice: A Good-Based Model. *Annual Review of Neuroscience* **34**, 333–359 (2011).

[6] Ratcliff, R. A theory of memory retrieval. *Psychological Review* **85**, 59–108 (1978).

[7] Orsolic, I., Rio, M., Mrsic-Flogel, T. D. & Znamenskiy, P. Mesoscale cortical dynamics reflect the interaction of sensory evidence and temporal expectation during perceptual decision-making. *Neuron* **109**, 1861–1875.e10 (2021).

[8] Hanks, T. D. *et al.* Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature* **520**, 220–223 (2015).

[9] Khilkevich, A. *et al.* Brain-wide dynamics linking sensation to action during decision-making. *Nature* **634**, 890–900 (2024).

[10] Kahneman, D. & Tversky, A. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* **47**, 263 (1979).

[11] Platt, M. L. & Glimcher, P. W. Neural correlates of decision variables in parietal cortex. *Nature* **400**, 233–238 (1999).

[12] Padoa-Schioppa, C. & Assad, J. A. Neurons in the orbitofrontal cortex encode economic value. *Nature* **441**, 223–226 (2006).

[13] Cai, X. & Padoa-Schioppa, C. Neuronal evidence for good-based economic decisions under variable action costs. *Nature Communications* **10**, 393 (2019).

[14] Rabin, M. & Thaler, R. H. Anomalies: Risk Aversion. *Journal of Economic Perspectives* **15**, 219–232 (2001).

[15] Giraldeau, L.-A. & Caraco, T. *Social Foraging Theory* Monographs in Behavior and Ecology (Princeton University Press, Princeton, New Jersey, 2000).

[16] Tervo, D. G. R. *et al.* Behavioral Variability through Stochastic Choice and Its Gating by Anterior Cingulate Cortex. *Cell* **159**, 21–32 (2014).

[17] Haroush, K. & Williams, Z. M. Neuronal Prediction of Opponent's Behavior during Cooperative Social Interchange in Primates. *Cell* **160**, 1233–1245 (2015).

[18] Ong, W. S., Madlon-Kay, S. & Platt, M. L. Neuronal correlates of strategic cooperation in monkeys. *Nature Neuroscience* **24**, 116–128 (2021).

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

[20] Dorris, M. C. & Glimcher, P. W. Activity in Posterior Parietal Cortex Is Correlated with the Relative Subjective Desirability of Action. *Neuron* **44**, 365–378 (2004).

[21] Von Neumann, J. & Morgenstern, O. *Theory of Games and Economic Behavior* 3rd ed edn (Princeton University press, Princeton (N.J.), 1953).

[22] Camerer, C. *Behavioral Game Theory: Experiments in Strategic Interaction* The Roundtable Series in Behavioral Economics (Russell Sage Foundation ; Princeton University Press, New York, N.Y. : Princeton, N.J, 2003).

[23] Nash, J. F. Equilibrium points in n -person games. *Proceedings of the National Academy of Sciences* **36**, 48–49 (1950).

[24] Hoy, J. L., Yavorska, I., Wehr, M. & Niell, C. M. Vision Drives Accurate Approach Behavior during Prey Capture in Laboratory Mice. *Current Biology* **26**, 3046–3052 (2016).

[25] Lin, H.-T. & Leonardo, A. Heuristic Rules Underlying Dragonfly Prey Selection and Interception. *Current Biology* **27**, 1124–1137 (2017).

[26] Coen, P. *et al.* Dynamic sensory cues shape song structure in Drosophila. *Nature* **507**, 233–237 (2014).

[27] Zilkha, N. *et al.* Sex-dependent control of pheromones on social organization within groups of wild house mice. *Current Biology* **33**, 1407–1420.e4 (2023).

[28] Neunuebel, J. P., Taylor, A. L., Arthur, B. J. & Egnor, S. R. Female mice ultrasonically interact with males during courtship displays. *eLife* **4**, e06203 (2015).

[29] Chan, T. C. Y., Fearing, D. S., Fernandes, C. & Kovalchik, S. A Markov process approach to untangling intention versus execution in tennis. *Journal of Quantitative Analysis in Sports* **18**, 127–145 (2022).

[30] Marcelino, R. *et al.* Collective movement analysis reveals coordination tactics of team players in football matches. *Chaos, Solitons & Fractals* **138**, 109831 (2020).

[31] Sweis, B. M. *et al.* Sensitivity to "sunk costs" in mice, rats, and humans. *Science* **361**, 178–181 (2018).

[32] Ormond, J. & O'Keefe, J. Hippocampal place cells have goal-oriented vector fields during navigation. *Nature* **607**, 741–746 (2022).

[33] El-Gaby, M. *et al.* A cellular basis for mapping behavioural structure. *Nature* **636**, 671–680 (2024).

[34] Nagy, M. *et al.* Synergistic Benefits of Group Search in Rats. *Current Biology* **30**, 4733–4738.e4 (2020).

[35] Rosenberg, M., Zhang, T., Perona, P. & Meister, M. Mice in a labyrinth show rapid learning, sudden insight, and efficient exploration. *eLife* **10**, e66175 (2021).

[36] Pinto, L. *et al.* Task-Dependent Changes in the Large-Scale Dynamics and Necessity of Cortical Regions. *Neuron* **104**, 810–824.e9 (2019).

[37] Aronov, D. & Tank, D. W. Engagement of Neural Circuits Underlying 2D Spatial Navigation in a Rodent Virtual Reality System. *Neuron* **84**, 442–456 (2014).

[38] Simon, P., Dupuis, R. & Costentin, J. Thigmotaxis as an index of anxiety in mice. Influence of dopaminergic transmissions. *Behavioural Brain Research* **61**, 59–64 (1994).

[39] Heitz, R. P. The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience* **8** (2014).

[40] Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).

[41] Machens, C. K. Flexible Control of Mutual Inhibition: A Neural Model of Two-Interval Discrimination. *Science* **307**, 1121–1124 (2005).

[42] Kopec, C. D., Erlich, J. C., Brunton, B. W., Deisseroth, K. & Brody, C. D. Cortical and Subcortical Contributions to Short- Term Memory for Orienting Movements. *Neuron* **88**, 367–377 (2015).

[43] Wong, K.-F. & Wang, X.-J. A Recurrent Network Mechanism of Time Integration in Perceptual Decisions. *Journal of Neuroscience* **26**, 1314–1328 (2006).

[44] Sridhar, V. H. *et al.* The geometry of decision-making in individuals and collectives. *Proceedings of the National Academy of Sciences* **118**, e2102157118 (2021).

[45] Shamash, P. *et al.* Mice learn multi-step routes by memorizing subgoal locations. *Nature Neuroscience* **24**, 1270–1279 (2021).

[46] Holmgren, C. D. *et al.* Visual pursuit behavior in mice maintains the pursued prey on the retinal region with least optic flow. *eLife* **10**, e70838 (2021).

[47] Euston, D. R. & McNaughton, B. L. Apparent Encoding of Sequential Context in Rat Medial Prefrontal Cortex Is Accounted for by Behavioral Variability. *Journal of Neuroscience* **26**, 13143–13155 (2006).

[48] Padilla-Coreano, N. *et al.* Cortical ensembles orchestrate social competition through hypothalamic outputs. *Nature* **603**, 667–671 (2022).

[49] Li, S. W. *et al.* Frontal neurons driving competitive behaviour and ecology of social groups. *Nature* **603**, 661–666 (2022).

[50] Zhou, T. *et al.* History of winning remodels thalamo-PFC circuit to reinforce social dominance. *Science* **357**, 162–168 (2017).

[51] Kingsbury, L. *et al.* Correlated Neural Activity and Encoding of Behavior across Brains of Socially Interacting Animals. *Cell* **178**, 429–446.e16 (2019).

[52] Vinyals, O. *et al.* Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575**, 350–354 (2019).

[53] Piet, A. T., Erlich, J. C., Kopec, C. D. & Brody, C. D. Rat Prefrontal Cortex Inactivations during Decision Making Are Explained by Bistable Attractor Dynamics. *Neural Computation* **29**, 2861–2886 (2017).

[54] Bao, C. *et al.* The rat frontal orienting field dynamically encodes value for economic decisions under risk. *Nature Neuroscience* **26**, 1942–1952 (2023).

[55] Chaturvedi, S., El-Gazzar, A. & Van Gerven, M. A dynamical systems approach to optimal foraging. *PLOS Complex Systems* **1**, e0000018 (2024).

[56] Liu, M., Nair, A., Coria, N., Linderman, S. W. & Anderson, D. J. Encoding of female mating dynamics by a hypothalamic line attractor. *Nature* **634**, 901–909 (2024).

[57] Gorbonos, D., Gov, N. S. & Couzin, I. D. Geometrical Structure of Bifurcations during Spatial Decision-Making. *PRX Life* **2**, 013008 (2024).

[58] Nair, A. *et al.* An approximate line attractor in the hypothalamus encodes an aggressive state. *Cell* **186**, 178–193.e15 (2023).

[59] Scherbaum, S. *et al.* Process dynamics in delay discounting decisions: An attractor dynamics approach. *Judgment and Decision Making* **11**, 472–495 (2016).

[60] Laibson, D. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics* **112**, 443–478 (1997).

[61] Gangopadhyay, P., Chawla, M., Dal Monte, O. & Chang, S. W. C. Prefrontal–amygdala circuits in social decision-making. *Nature Neuroscience* **24**, 5–18 (2021).

[62] Hennequin, G., Vogels, T. P. & Gerstner, W. Optimal Control of Transient Dynamics in Balanced Networks Supports Generation of Complex Movements. *Neuron* **82**, 1394–1406 (2014).

[63] Smith, K. Lab mice go wild: Making experiments more natural in order to decode the brain. *Nature* **618**, 448–450 (2023).

[64] Dennis, E. J. *et al.* Systems Neuroscience of Natural Behaviors in Rodents. *The Journal of Neuroscience* **41**, 911–919 (2021).

[65] Pereira, T. D. *et al.* SLEAP: A deep learning system for multi-animal pose tracking. *Nature Methods* **19**, 486–495 (2022).

[66] Lopes, G. *et al.* Bonsai: An event-based framework for processing and controlling data streams. *Frontiers in Neuroinformatics* **9** (2015).

[67] Buccino, A. P. *et al.* Open source modules for tracking animal behavior and closed-loop stimulation based on Open Ephys and Bonsai. *Journal of Neural Engineering* **15**, 055002 (2018).

[68] Lauer, J. *et al.* Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nature Methods* **19**, 496–504 (2022).

[69] Walter, T. & Couzin, I. D. TRex, a fast multi-animal tracking system with markerless identification, and 2D estimation of posture and visual fields. *eLife* **10**, e64000 (2021).

[70] Mobbs, D., Trimmer, P. C., Blumstein, D. T. & Dayan, P. Foraging for foundations in decision neuroscience: Insights from ethology. *Nature Reviews Neuroscience* **19**, 419–427 (2018).

[71] Lockwood, P. L., Apps, M. A. & Chang, S. W. Is There a 'Social' Brain? Implementations and Algorithms. *Trends in Cognitive Sciences* **24**, 802–813 (2020).

[72] Testard, C. *et al.* Neural signatures of natural behaviour in socializing macaques. *Nature* **628**, 381–390 (2024).

[73] Forli, A. & Yartsev, M. M. Hippocampal representation during collective spatial behaviour in bats. *Nature* **621**, 796–803 (2023).

[74] Ray, S. *et al.* Hippocampal coding of identity, sex, hierarchy, and affiliation in a social group of wild fruit bats. *Science* **387**, eadk9385 (2025).

[75] Furtak, S. C. *et al.* The Floor Projection Maze: A novel behavioral apparatus for presenting visual stimuli to rats. *Journal of Neuroscience Methods* **181**, 82–88 (2009).

[76] Sainsbury Wellcome Centre Foraging Behaviour Working Group. Aeon: An open-source platform to study the neural basis of ethological behaviours over naturalistic timescales. Zenodo (2023).

[77] Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. Julia: A Fresh Approach to Numerical Computing. *SIAM Review* **59**, 65–98 (2017).

[78] Alday, P. M. & Bates, D. MixedModels.jl. Zenodo (2025).

[79] Danisch, S. & Krumbiegel, J. Makie.jl: Flexible high-performance data visualization for Julia. *Journal of Open Source Software* **6**, 3349 (2021).

[80] Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)* **9**, 249–256 (2010).

[81] Rackauckas, C. & Nie, Q. DifferentialEquations.jl–a performant and feature-rich ecosystem for solving differential equations in Julia. *Journal of Open Research Software* **5** (2017).

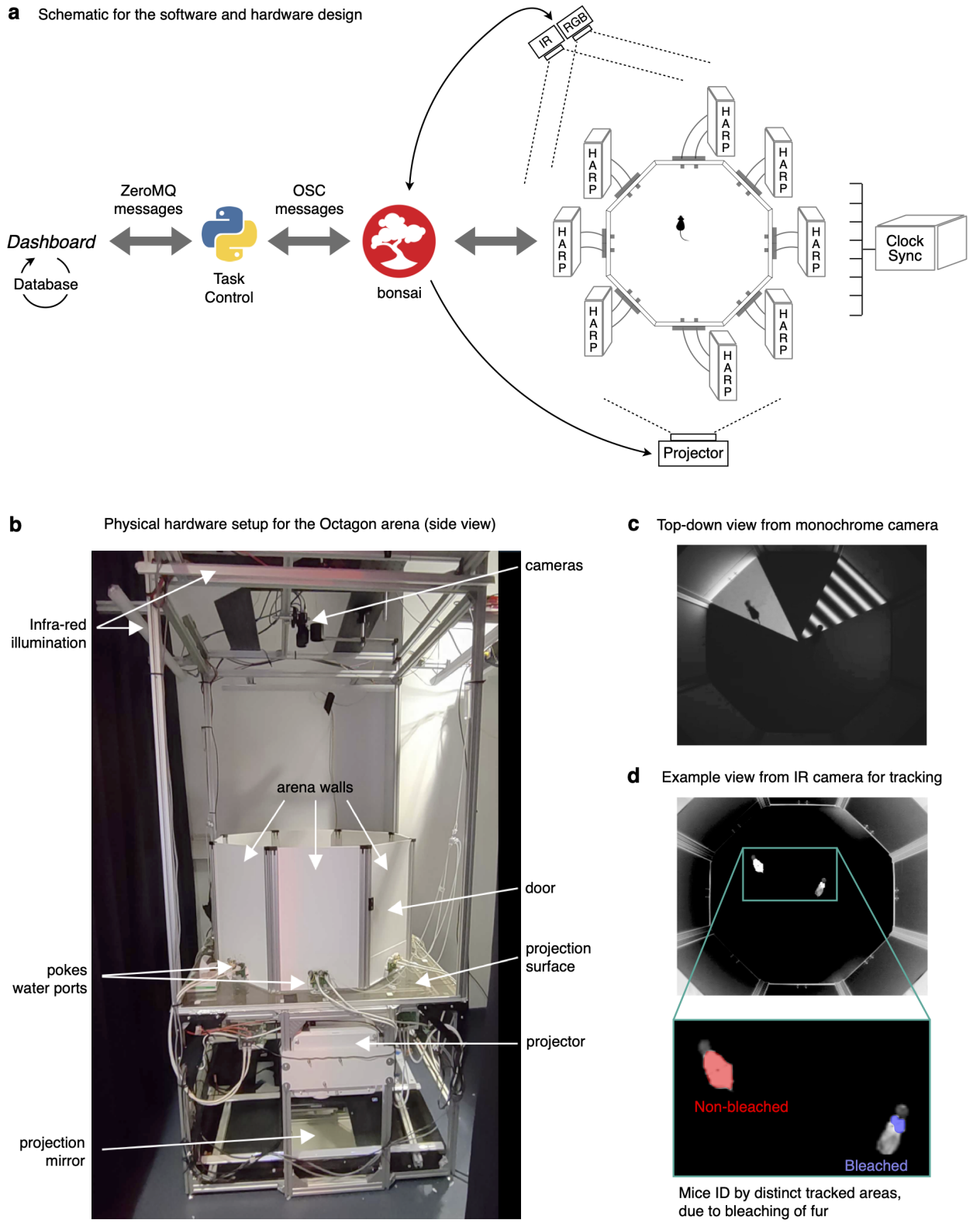

Fig. S1: Design of the Octagon arena. **a**, Schematic for software (Bonsai, Python) and hardware (HARP) design, task control, data acquisition and storage. **b**, Side view of the Octagon setup. **c**, Example top-down view from the monochrome camera, where both the visual stimuli and mice could be monitored. **d**, Example view (same frame as **c**) from the IR camera, which was used for position and identification tracking. Note that the tracked area size for the non-bleached mouse was larger than that of the bleached mouse (Methods).

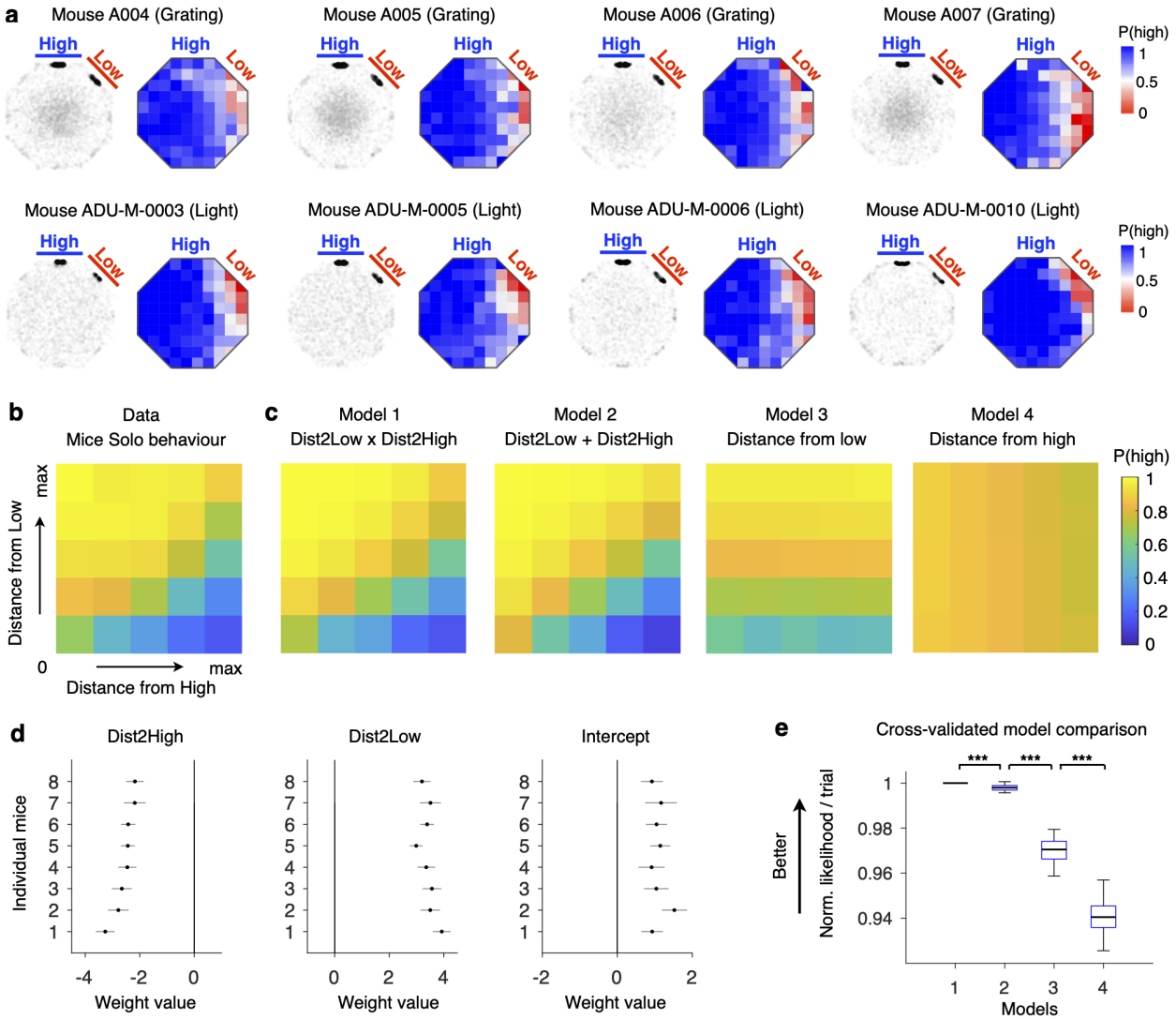

Fig. S2: Solo preference for all individual mice. **a**, Tracked positions at stimulus onset (gray) and at the time of response (black, left) and post-learning performance across varied start positions in the Octagon (right), rotated and reflected for visualization, for 8 individual mice. All mice had a preference for the high reward stimuli, generalized across start positions and sensory-value mappings (top 4 mice, grating was associated with high reward; bottom 4 mice, light was associated with high reward). **b,c**, Performance as a function of normalized distance to the high reward patch and the low reward patch, for mice data (**b**, 29007 trials from 235 sessions and 8 mice), and for predictions from different GLMM models (**c**). Model 3 (Table S3) and model 4 (Table S4) only included the distance from the low patch regressor and the distance from the high patch regressor respectively; model 2 (Table S2) had both regressors as main effects; and model 1 (Table S4) had both regressors and their interaction. **d**, Regressor coefficients for the main effects of distance to the high patch; distance to the low patch; and the intercept from model 2, fit to each of the 8 individual mice. All mice had a general preference for the high patch, and their preference was modulated bidirectionally by the distances to the high and low patches. Error bars represent 99% confidence intervals. **e**, Normalized negative log likelihood per trial across different GLMM models, with 20 fold cross-validation. *** $P < 0.001$, paired bootstrapped tests. Including each distance regressor and their interaction significantly improved model performance.

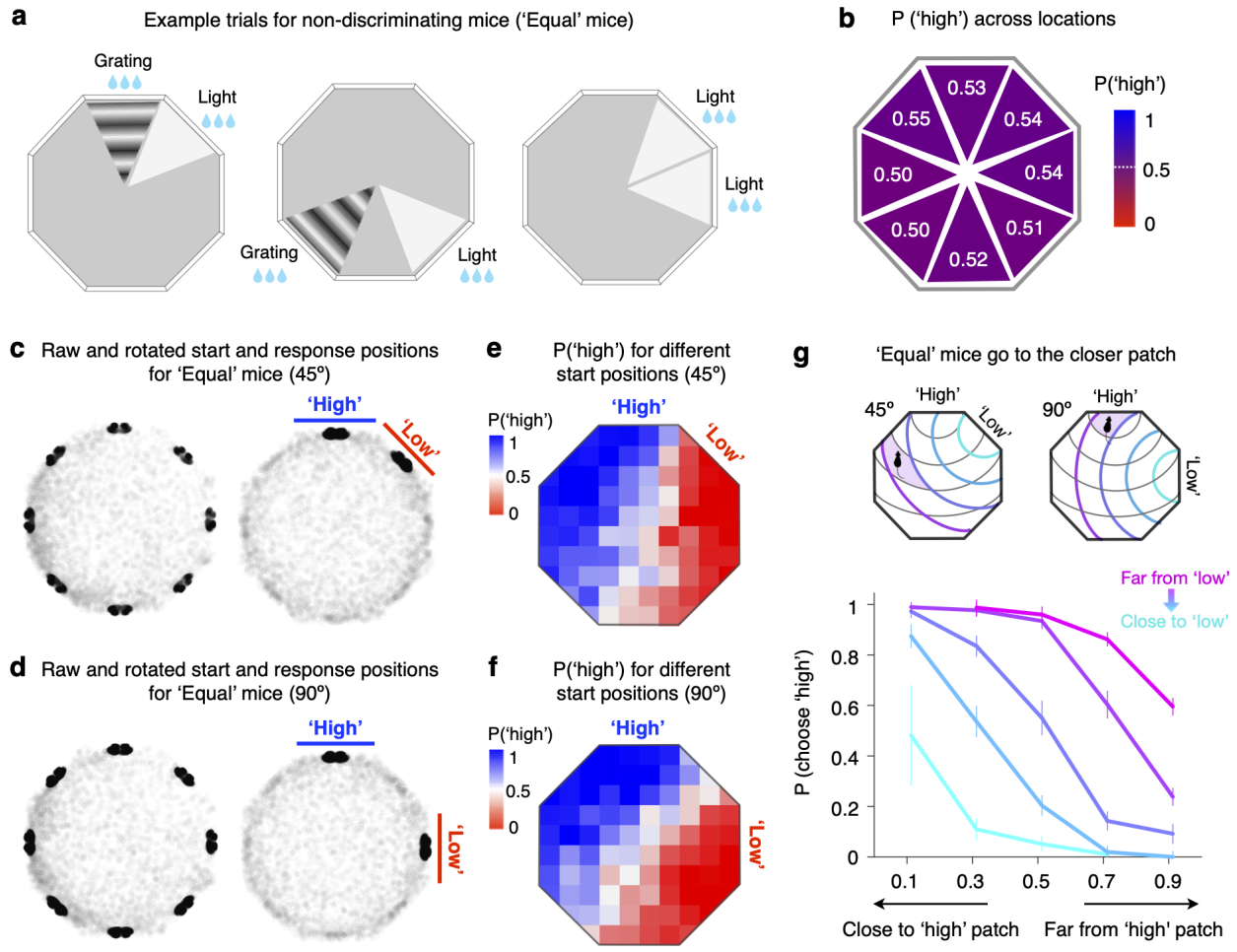

Fig. S3: Behaviours in control ('equal') mice. **a**, Example visual stimuli for 'equal mice', identical to those used in the experimental group, except that each stimulus was associated with the same amount of reward (3 units). **b**, P('high') is defined as the probability of choosing the light patch. 'Equal' mice did not develop a preference for any particular stimulus across different patch locations, in contrast to Fig. 1**h**. **c,d**, 'Equal' mice did not develop a strategy of clustering towards the arena centre at trial start, in contrast to Fig. 1**k,m**. **e-g**, 'Equal' mice made choices based on proximity, in contrast to the experimental group shown in Fig. 1**l,n,o**.

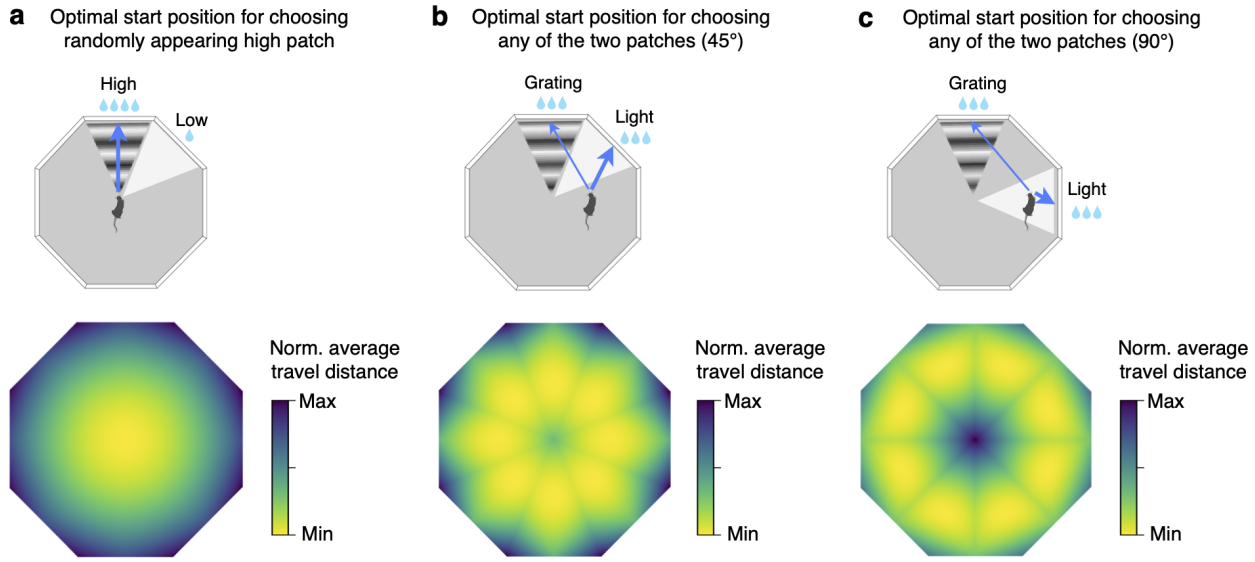

Fig. S4: Optimal start positions to minimize average travel distances for distinct goals. **a**, To minimize average travel distances for a randomly appearing patch, such as choosing the high reward patch no matter where it is, the optimal start position is in the center of the arena. **b,c**, Similar to **a**, but for goals of choosing the closer of the two randomly appearing patches, grating or light, when patches are 45 (**b**) or 90 (**c**) degrees apart. Note that start positions closer to edges can increase the probability of being close to one of the two patches, and is thus more optimal for a proximity-dominated strategy.

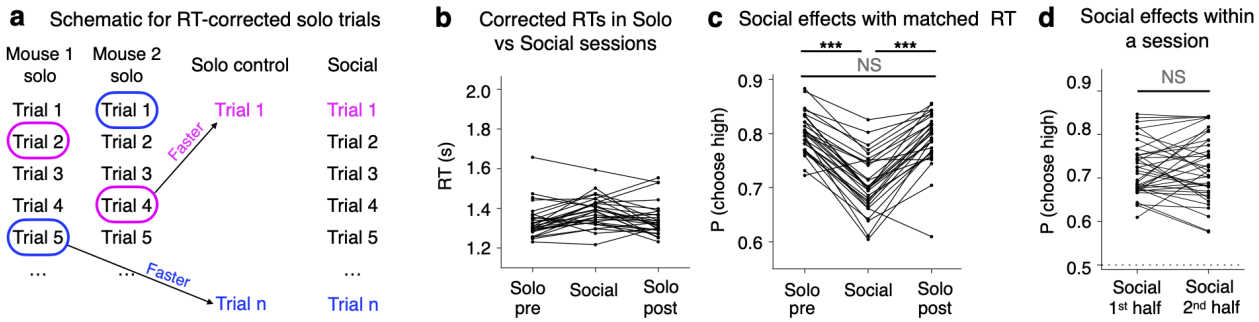

Fig. S5: Preference changes across contexts when controlled for response time (RT) and session time. **a**, Schematic for how RT-corrected solo control trials were selected. For each trial in the social session, a pair of trials were randomly chosen, with replacement, from two animals' solo control sessions 1 day before. To simulate race without competition, the faster trial from the pair was included as the corresponding solo control trial. **b**, Correction resulted in comparable RTs across contexts. **c**, With matched RT distributions, mice still decreased preference for the high reward patch under competition, similar to Fig. 2**c**. **d**, The first and second halves of social sessions showed comparable performance, arguing against a slowly changing preference due to non-social learning. ***, $P < 0.001$; NS, not significant, $P > 0.05$; paired t-tests.

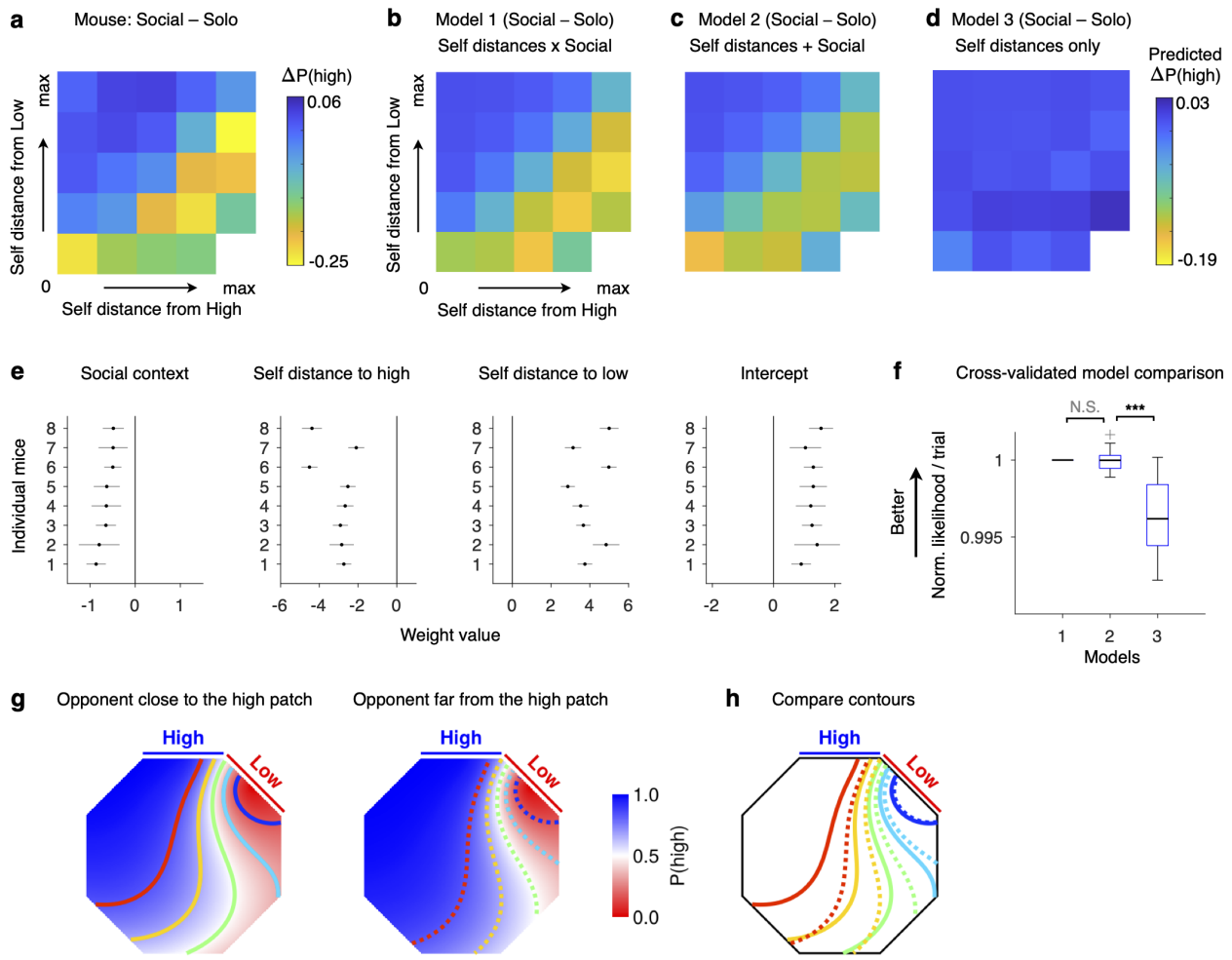

Fig. S6: GLMM fits for the effects of social competition on performance. **a**, Decrease in the probability of choosing the high reward patch in social sessions (7035 trials, 8 mice) compared to solo sessions (9091 trials, 8 mice), as a function of the chooser's distances to the high and low patches. **b-d**, Similar to **a**, for predictions from different GLMM models (Methods). Model 3 (Table S8) only included the distance regressors but not the social context; model 2 (Table S7) had both the distance regressors and the social context main effect; and model 1 (Table S6) included both distance regressors, social context main effect and their interactions. **e**, Regressor coefficients for the main effects of the social context; distance to the high patch; distance to the low patch; and the intercept, fit to each of the 8 individual mice. All mice had decreased preference for the high reward patch under social competition. Error bars represent 95% confidence intervals. **f**, Normalized negative log likelihood per trial across different GLMM models, with 20 fold cross-validation. ***, $P < 0.001$; NS, not significant $P = 0.85$, paired bootstrapped tests. Including the social context as a main effect significantly improved model performance. **g**, Visualization of the GLMM fit to the mouse behavior as a function of opponent distance to the high patch (Table S12). For the left panel `other2high = 0` and for the right panel `other2high = 2`. The red to blue color indicates P(high) at each self start position in the octagon. To aid comparison of the change in preference across the two panels, we drew contour lines at P(high)= $\frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}$ (in blue, cyan, green, yellow, red respectively), splitting the data into sextiles. To make this visualization clearer, we only simulated 45° trials. **h**, the contours from **g** replotted together for easier comparison. Note how the blue contour is relatively unchanged, and the difference between the contours grows as the P(high) grows, especially when self is far from the patches.

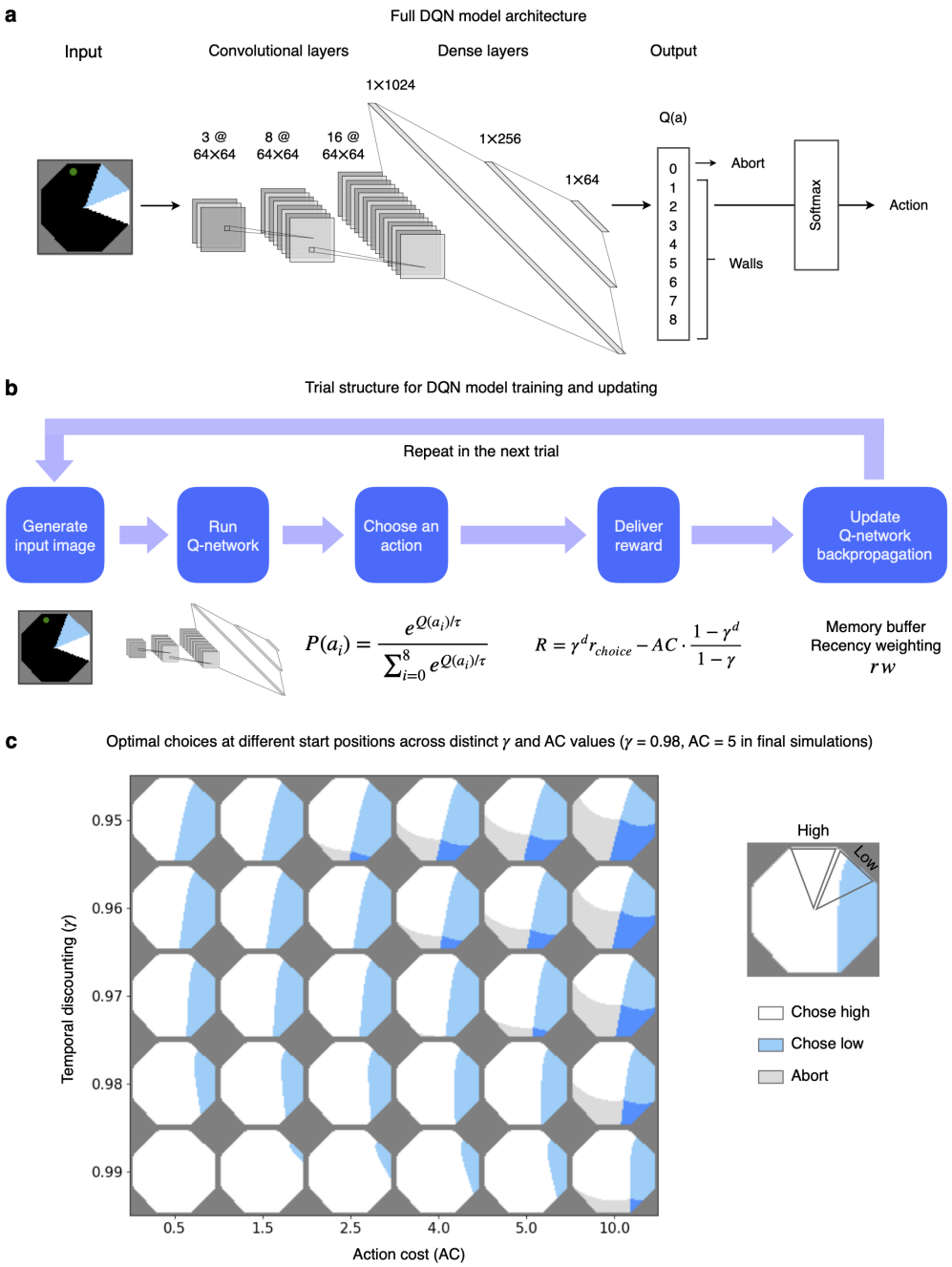

Fig. S7: DQN details. **a**, Detailed model architecture. **b**, Detailed trial structure for input, choice generation, discounted reward delivery and network updating (Methods). **c**, Simulations for optimal choices at different start positions in the Octagon with parametrically varied values of temporal discounting (γ) and action cost (AC). $\gamma = 0.98$ and AC $= 5$ were chosen to qualitatively match mice solo preference.

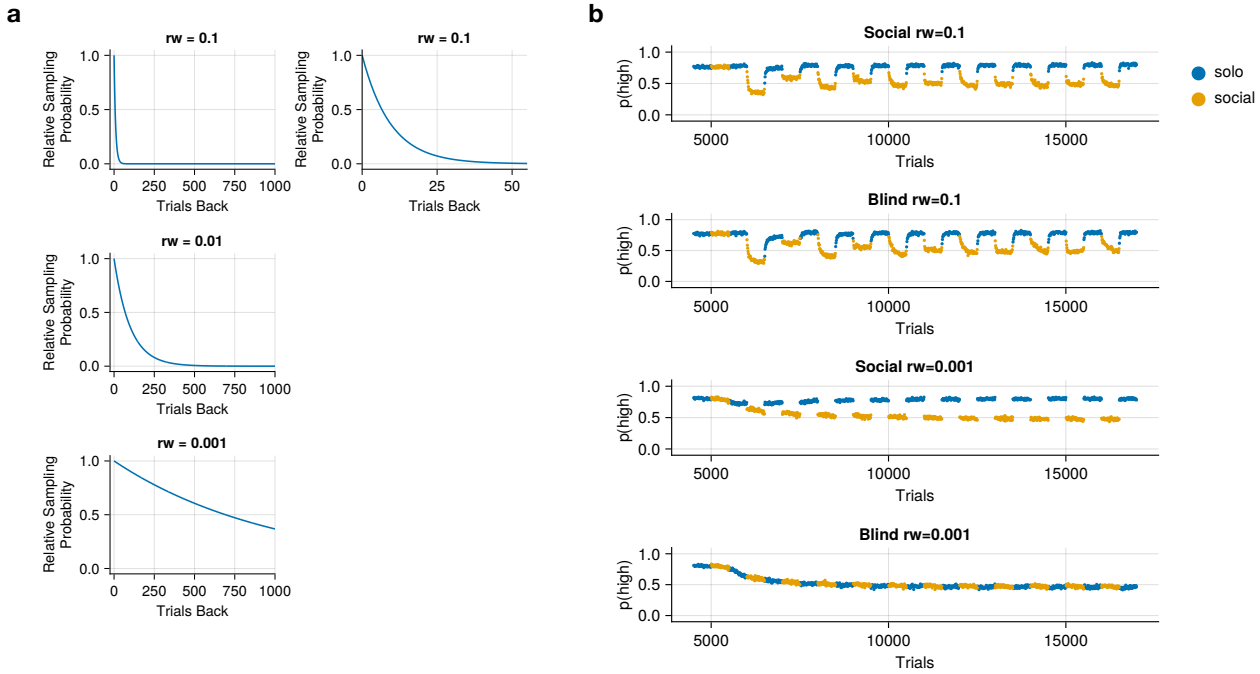

Fig. S8: Influence of recency weighting on DQN behavior. **a**, Visualization of the priority sampling for different levels of recency weighting (rw). Each row shows the sampling for a smaller rw. For rw=0.1, we show a zoomed in plot (top right) to show that samples are from very recent history (e.g. < 50 trials back). **b**, Each point is the average P(high) of 10 trials, averaged across all simulations. Trials from solo sessions are in blue and social sessions are in orange. For recency weighting (rw) of 0.1, both 'social' and 'socially-blind' agents switched strategies across blocks, but showed evidence of within block learning. In contrast, for rw=0.001, the 'social' agent had context-specific strategies, while the 'socially-blind' agent did not.

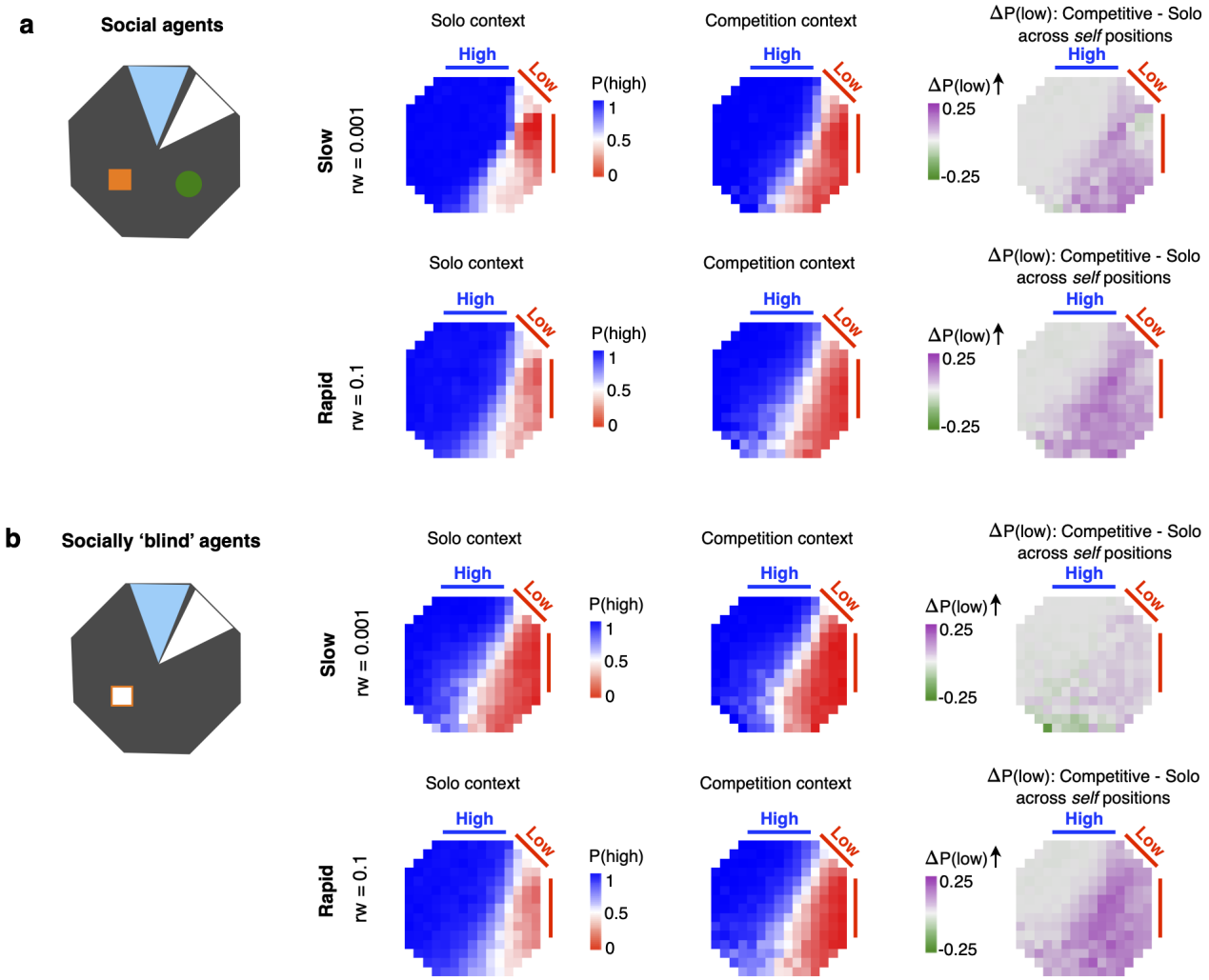

Fig. S9: 'Self'-position-specific preference shifts in competing DQN agents. **a**, For 'social' DQN agents, preference shifts across contexts varied as a function of 'self' positions in the arena, for both slow learning (top row, rw = 0.001) and rapid learning (bottom row, rw = 0.1). These patterns were similar to mice behaviours (Fig. 2**e,f**). **b**, Similar to **a**, for 'socially-blind' agents. Note that with rapid learning, even 'socially-blind' agents could display self-position-specific preference shifts.

Statistical Appendix

Mice Solo Models

For the GLMM exploring the factors influencing Solo decisions, the column definitions were as follows:

Column Name	Description
choose_high	Boolean variable, 1 if the mouse chose the high patch.
dist2high	Distance of the mouse to the high patch.
dist2low	Distance of the mouse to the low patch.
subjid	A categorical variable identifying the mouse

```
Model 1 (full solo model)
Formula: choose_high ~ 1 + dist2high*dist2low + (1 + dist2high + dist2low | subjid)

Model fit statistics:
    AIC           BIC           LogLikelihood    Deviance
    1.6321e+05    1.6329e+05    -81595           1.6319e+05

Fixed effects coefficients (95% CIs):
    Name                          Estimate    SE         tStat      DF       pValue
    {'(Intercept)'       }        -0.36269    0.13422    -2.7022    28979      0.0068922
    {'dist2high'         }         -1.2268    0.13778    -8.9046    28979     5.6625e-19
    {'dist2low'          }          5.1992    0.16422      31.66    28979    2.6614e-216
    {'dist2high:dist2low'}         -1.4618    0.11013    -13.274    28979     4.3055e-40
```

Table S1: Mice Solo Model 1

```
Model 2 (high and low)
Formula: choose_high ~ 1 + dist2high + dist2low + (1 + dist2high + dist2low | subjid)

Model fit statistics:
    AIC           BIC           LogLikelihood    Deviance
    1.5669e+05    1.5677e+05    -78337           1.5667e+05

Fixed effects coefficients (95% CIs):
    Name                   Estimate    SE          tStat      DF       pValue
    {'(Intercept)'}         1.0937     0.064651     16.918    28980     6.7781e-64
    {'dist2high'  }        -2.5334      0.11439    -22.148    28980     8.653e-108
    {'dist2low'   }         3.4084     0.096151     35.448    28980    1.7588e-269
```

Table S2: Mice Solo Model 2

```
Model 3 (low alone)
Formula: choose_high ~ 1 + dist2low + (1 + dist2low | subjid)

Model fit statistics:
    AIC           BIC           LogLikelihood    Deviance
    1.4924e+05    1.4928e+05    -74615           1.4923e+05

Fixed effects coefficients (95% CIs):
    Name                   Estimate    SE          tStat     DF       pValue
    {'(Intercept)'}        -0.52732    0.093084    -5.665    28981     1.4843e-08
    {'dist2low'   }          2.2228    0.076099    29.209    28981    7.1486e-185
```

Table S3: Mice Solo Model 3

```
Model 4 (high alone)
Formula: choose_high ~ 1 + dist2high + (1 + dist2high | subjid)
```

```
Model fit statistics:
    AIC           BIC           LogLikelihood    Deviance
    1.3909e+05    1.3913e+05    -69541           1.3908e+05

Fixed effects coefficients (95% CIs):
    Name                   Estimate    SE          tStat      DF       pValue
    {'(Intercept)'}           2.227    0.073248     30.404    28981    6.7323e-200
    {'dist2high'  }        -0.61688    0.074876    -8.2387    28981     1.8142e-16
```

Table S4: Mice Solo Model 3

```
20-fold cross validation, neg_LLH/trial, then bootmean for stats across folds for each model comparison

Model 1    Model 2    Model 3    Model4
-------    -------    -------    -------

0.68747    0.68764    0.66386    0.62337
0.69352    0.69417    0.67295    0.63591
0.69376    0.69186    0.66246    0.62958
0.68958    0.68681    0.65577    0.61514
0.69641    0.69323    0.66393    0.63058
0.68842    0.68745    0.6586     0.63648
0.69521    0.69269    0.6618     0.64245
0.68794    0.68513    0.66274    0.63274
0.68224    0.68009    0.64097    0.61815
0.71759    0.71587    0.68171    0.65122
0.67601    0.67627    0.65432    0.63298
0.71415    0.71306    0.67668    0.64475
0.68197    0.67893    0.64759    0.6224
0.6863     0.68284    0.65714    0.62871
0.68144    0.68041    0.65953    0.62751
0.70249    0.70066    0.66874    0.63987
0.6829     0.68154    0.65518    0.62226
0.68909    0.68478    0.66268    0.6301
0.7031     0.69975    0.67622    0.64364
0.67442    0.67095    0.64704    0.62454
```

Table S5: Mice Solo 20-fold cross validation

Mice Solo vs. Social (without opponent position)

For the GLMM exploring the overall effect of solo vs. social context, the column definitions were as follows:

Column Name	Description
choose_high	Boolean variable 1 if the 'self' mouse chose the high patch.
dist2high	Distance of the 'self' mouse to the high patch.
dist2low	Distance of the 'self' mouse to the low patch.
Social	Boolean variable. 1 if the session was a social session.
subjid	A categorical variable identifying the mouse

The notation for the formula follows Wilkinson notation as used by the `fitglme` function in MATLAB.

```
Formula:
    choose_high ~ 1 + dist2high*dist2low + dist2high*Social + dist2low*Social + dist2high:dist2low:Social +
            (1 + dist2high + dist2low + Social | subjid)

Model fit statistics:
    AIC      BIC      LogLikelihood    Deviance
    87619    87757    -43791           87583
```

```
Fixed effects coefficients (95% CIs):
    Name                                Estimate    SE         tStat      DF       pValue
    {'(Intercept)'              }       0.009119    0.22096    0.04127    16039       0.96708
    {'dist2high'                }        -1.7641    0.32866    -5.3677    16039     8.087e-08
    {'dist2low'                 }         5.1945    0.35984     14.436    16039    6.0669e-47
    {'Social'                   }       -0.35355    0.28848    -1.2256    16039       0.22038
    {'dist2high:dist2low'       }        -1.2177    0.19317    -6.3039    16039    2.9775e-10
    {'dist2high:Social'         }       -0.50623    0.28152    -1.7982    16039      0.072167
    {'dist2low:Social'          }        0.23784    0.36422    0.65302    16039       0.51375
    {'dist2high:dist2low:Social'}       0.087511    0.27407     0.3193    16039        0.7495
```

Table S6: Solo vs. Social: Model 1

```
Formula:
    choose_high ~ 1 + Social + dist2high*dist2low + (1 + dist2high + dist2low + Social | subjid)

Model fit statistics:
    AIC      BIC      LogLikelihood    Deviance
    88277    88392    -44123           88247

Fixed effects coefficients (95% CIs):
    Name                          Estimate    SE          tStat      DF       pValue
    {'(Intercept)'       }         0.15636     0.16171    0.96693    16042        0.3336
    {'dist2high'         }          -2.035     0.29651    -6.8633    16042     6.976e-12
    {'dist2low'          }          5.3199     0.30863     17.237    16042     5.472e-66
    {'Social'            }        -0.61662    0.057115    -10.796    16042    4.4463e-27
    {'dist2high:dist2low'}         -1.1709     0.13761    -8.5089    16042    1.9089e-17
```

Table S7: Solo vs. Social: Model 2

```
Formula:
    choose_high ~ 1 + dist2high*dist2low + (1 + dist2high + dist2low | subjid)

Model fit statistics:
    AIC      BIC      LogLikelihood    Deviance
    87642    87719    -43811           87622

Fixed effects coefficients (95% CIs):
    Name                          Estimate    SE         tStat      DF       pValue
    {'(Intercept)'       }        -0.32374    0.15849    -2.0426    16043      0.041105
    {'dist2high'         }         -1.9003    0.30259    -6.2801    16043    3.4704e-10
    {'dist2low'          }          5.5203    0.30848     17.895    16043    6.2738e-71
    {'dist2high:dist2low'}         -1.2867    0.13648    -9.4275    16043    4.7602e-21
```

Table S8: Solo vs. Social: Model 3

```
20-fold cross validation, neg_LLH/trial, then bootmean for stats across folds for each model comparison

Model 1     Model 2   Model 3
-------    -------    -------

0.67278    0.67356    0.66691
0.68699    0.68588    0.68369
0.70654    0.70603    0.70036
0.6898     0.68973    0.68583
0.68479    0.68486    0.68291
0.68008    0.67956    0.67877
0.69653    0.69648    0.68873
0.64793    0.64737    0.64193
0.70412    0.70404    0.69887
0.68933    0.68965    0.68933
0.68044    0.67968    0.67951
0.66509    0.66422    0.66287
0.67773    0.67801    0.6753
0.6954     0.6965     0.69443
0.69276    0.69442    0.68589
0.66306    0.66319    0.65891
0.67933    0.67935    0.67951
0.68029    0.67927    0.67551
0.68078    0.68154    0.67687
0.66734    0.66743    0.66362
```

Table S9: Solo vs. Social: 20-fold cross validation

Opponent Effects GLMM

For the GLMM testing the effects of opponent position, the column definitions were as follows:

Column Name	Description
p1_chose_high	Boolean variable, 1 if player 1 chose the high patch.
p1_dh	Distance of player 1, 'self', to high patch.
p1_dl	Distance of player 1 to low patch.
p2_dh	Distance of player 2, 'other', to high patch.
p1_name	A categorical variable identifying the player 1 mouse (or DQN seed)

The notation for the formula follows Wilkinson notation as used by the `MixedModels.jl` package in Julia.

Mice

```
Generalized Linear Mixed Model fit by maximum likelihood (nAGQ = 1)
  p1_chose_high ~ 1 + p1_dh + p1_dl + p1_dh & p1_dl + (1 + p1_dh + p1_dl | p1_name)
  Distribution: Distributions.Bernoulli{Float64}
  Link: GLM.LogitLink()

   logLik    deviance     AIC       AICc        BIC
 -4470.2991  8940.5983  8960.5983  8960.6194  9033.1514

 Number of obs: 10460; levels of grouping factors: 8

Fixed-effects parameters:
-----------------------------------------------------
                   Coef.  Std. Error      z  Pr(>|z|)
-----------------------------------------------------
(Intercept)    -0.537957    0.208488  -2.58    0.0099
p1_dh          -1.89595     0.292589  -6.48    <1e-10
p1_dl           5.25532     0.289949  18.12    <1e-72
p1_dh & p1_dl  -1.22726     0.143604  -8.55    <1e-16
```

```
-----------------------------------------------------
```

Table S10: Mice: Model 1

```
Generalized Linear Mixed Model fit by maximum likelihood (nAGQ = 1)
  p1_chose_high ~ 1 + p1_dh + p1_dl + p2_dh + p1_dh & p1_dl + (1 + p1_dh + p1_dl | p1_name)
  Distribution: Distributions.Bernoulli{Float64}
  Link: GLM.LogitLink()

   logLik    deviance     AIC       AICc        BIC
 -4458.5569  8917.1137  8939.1137  8939.1390  9018.9222

 Number of obs: 10460; levels of grouping factors: 8

Fixed-effects parameters:
-----------------------------------------------------
                   Coef.  Std. Error      z  Pr(>|z|)
-----------------------------------------------------
(Intercept)    -0.84789    0.216339   -3.92    <1e-04
p1_dh          -1.90342    0.295777   -6.44    <1e-09
p1_dl           5.26813    0.293679   17.94    <1e-71
p2_dh           0.267875   0.0552366   4.85    <1e-05
p1_dh & p1_dl  -1.23615    0.143981   -8.59    <1e-17
-----------------------------------------------------
```

Table S11: Mice: Model 2

```
Generalized Linear Mixed Model fit by maximum likelihood (nAGQ = 1)
  p1_chose_high ~ 1 + p1_dh + p1_dl + p2_dh + p1_dh & p1_dl + p2_dh & p1_dh + p2_dh & p1_dl +
        (1 + p1_dh + p1_dl | p1_name)
  Distribution: Distributions.Bernoulli{Float64}
  Link: GLM.LogitLink()

   logLik    deviance     AIC       AICc        BIC
 -4452.7833  8905.5667  8931.5667  8931.6015  9025.8857

 Number of obs: 10460; levels of grouping factors: 8

Fixed-effects parameters:
-----------------------------------------------------
                   Coef.  Std. Error      z  Pr(>|z|)
-----------------------------------------------------
(Intercept)    -0.848081    0.266073  -3.19    0.0014
p1_dh          -1.36025     0.354533  -3.84    0.0001
p1_dl           4.65406     0.342752  13.58    <1e-41
p2_dh           0.253759    0.153136   1.66    0.0975
p1_dh & p1_dl  -1.2611      0.144605  -8.72    <1e-17
p2_dh & p1_dh  -0.455316    0.17106   -2.66    0.0078
p2_dh & p1_dl   0.560817    0.16528    3.39    0.0007
-----------------------------------------------------
```

Table S12: Mice: Model 3

```
Likelihood-ratio test: 3 models fitted on 10460 observations
-------------------------------------------------------
     DOF  dDOF      LogLik   Deviance    Chisq  p(>Chisq)
-------------------------------------------------------
[1]   10        -4470.2991  8940.5983
[2]   11     1  -4458.5569  8917.1137  23.4846     <1e-05
[3]   13     2  -4452.7833  8905.5667  11.5471     0.0031
-------------------------------------------------------
```

Table S13: Mice, likelihood ratio tests.

Social RL

```
Generalized Linear Mixed Model fit by maximum likelihood (nAGQ = 1)
  p1_chose_high ~ 1 + p1_dh + p1_dl + p1_dh & p1_dl + (1 + p1_dh + p1_dl | p1_name)
  Distribution: Distributions.Bernoulli{Float64}
  Link: GLM.LogitLink()

   logLik    deviance     AIC       AICc        BIC
 -4519.0709  9038.1418  9058.1418  9058.1638  9130.2452

 Number of obs: 10000; levels of grouping factors: 10

Fixed-effects parameters:
-----------------------------------------------------
                   Coef.  Std. Error       z  Pr(>|z|)
-----------------------------------------------------
(Intercept)     2.68442     0.212669   12.62    <1e-35
p1_dh          -3.77314     0.19803   -19.05    <1e-80
p1_dl           1.8412      0.191092    9.64    <1e-21
p1_dh & p1_dl  -0.396122    0.155857   -2.54    0.0110
-----------------------------------------------------
```

Table S14: Social RL Model 1

```
Generalized Linear Mixed Model fit by maximum likelihood (nAGQ = 1)
  p1_chose_high ~ 1 + p1_dh + p1_dl + p2_dh + p1_dh & p1_dl + (1 + p1_dh + p1_dl | p1_name)
  Distribution: Distributions.Bernoulli{Float64}
  Link: GLM.LogitLink()

   logLik    deviance     AIC       AICc        BIC
 -3902.6969  7805.3938  7827.3938  7827.4202  7906.7075

 Number of obs: 10000; levels of grouping factors: 10

Fixed-effects parameters:
-----------------------------------------------------
                   Coef.  Std. Error       z  Pr(>|z|)
-----------------------------------------------------
(Intercept)     0.86314    0.242954     3.55    0.0004
p1_dh          -4.57338    0.219619   -20.82    <1e-95
p1_dl           1.99823    0.20626      9.69    <1e-21
p2_dh           2.09977    0.0667218   31.47    <1e-99
p1_dh & p1_dl  -0.304157   0.166302    -1.83    0.0674
-----------------------------------------------------
```

Table S15: Social RL Model 2

```
Generalized Linear Mixed Model fit by maximum likelihood (nAGQ = 1)
  p1_chose_high ~ 1 + p1_dh + p1_dl + p2_dh + p1_dh & p1_dl + p2_dh & p1_dh + p2_dh & p1_dl +
        (1 + p1_dh + p1_dl | p1_name)
  Distribution: Distributions.Bernoulli{Float64}
  Link: GLM.LogitLink()

   logLik    deviance     AIC       AICc        BIC
 -3852.0240  7704.0480  7730.0480  7730.0844  7823.7824

 Number of obs: 10000; levels of grouping factors: 10

Fixed-effects parameters:
```

```
-----------------------------------------------------
                   Coef.  Std. Error       z  Pr(>|z|)
-----------------------------------------------------
(Intercept)     2.13966     0.30458     7.02    <1e-11
p1_dh          -4.36593     0.318756  -13.70    <1e-41
p1_dl           1.31971     0.231426    5.70    <1e-07
p2_dh           0.411138    0.20667     1.99    0.0467
p1_dh & p1_dl  -0.935697    0.17934    -5.22    <1e-06
p2_dh & p1_dh   0.387703    0.189478    2.05    0.0407
p2_dh & p1_dl   1.17338     0.149839    7.83    <1e-14
-----------------------------------------------------
```

Table S16: Social RL Model 3

```
Likelihood-ratio test: 3 models fitted on 10000 observations
-------------------------------------------------------------
     DOF  dDOF      LogLik   Deviance      Chisq  p(>Chisq)
-------------------------------------------------------------
[1]   10        -4519.0709  9038.1418
[2]   11     1  -3902.6969  7805.3938  1232.7480     <1e-99
[3]   13     2  -3852.0240  7704.0480   101.3458     <1e-22
-------------------------------------------------------------
```

Table S17: Social RL, likelihood ratio tests.

Socially-Blind RL

```
Generalized Linear Mixed Model fit by maximum likelihood (nAGQ = 1)
  p1_chose_high ~ 1 + p1_dh + p1_dl + p1_dh & p1_dl + (1 + p1_dh + p1_dl | p1_name)
  Distribution: Distributions.Bernoulli{Float64}
  Link: GLM.LogitLink()

   logLik    deviance     AIC       AICc        BIC
 -4684.0539  9368.1077  9388.1077  9388.1297  9460.2111

 Number of obs: 10000; levels of grouping factors: 10

Fixed-effects parameters:
-----------------------------------------------------
                   Coef.  Std. Error       z  Pr(>|z|)
-----------------------------------------------------
(Intercept)     2.45723      0.204452   12.02    <1e-32
p1_dh          -4.02843      0.189319  -21.28    <1e-99
p1_dl           2.04995      0.1891     10.84    <1e-26
p1_dh & p1_dl  -0.0151294    0.144481   -0.10    0.9166
-----------------------------------------------------
```

Table S18: Socially-blind RL Model 1

```
Generalized Linear Mixed Model fit by maximum likelihood (nAGQ = 1)
  p1_chose_high ~ 1 + p1_dh + p1_dl + p2_dh + p1_dh & p1_dl + (1 + p1_dh + p1_dl | p1_name)
  Distribution: Distributions.Bernoulli{Float64}
  Link: GLM.LogitLink()

   logLik    deviance     AIC       AICc        BIC
 -4683.9046  9367.8092  9389.8092  9389.8356  9469.1229

 Number of obs: 10000; levels of grouping factors: 10

Fixed-effects parameters:
-----------------------------------------------------
```

```
                    Coef.  Std. Error       z  Pr(>|z|)
-------------------------------------------------------
(Intercept)     2.48849     0.212582    11.71    <1e-30
p1_dh          -4.02778     0.189328   -21.27    <1e-99
p1_dl           2.05159     0.189202    10.84    <1e-26
p2_dh          -0.0278421   0.0522742   -0.53    0.5943
p1_dh & p1_dl  -0.0166      0.144526    -0.11    0.9086
-------------------------------------------------------
```

Table S19: Socially-blind RL Model 2

```
Generalized Linear Mixed Model fit by maximum likelihood (nAGQ = 1)
  p1_chose_high ~ 1 + p1_dh + p1_dl + p2_dh + p1_dh & p1_dl + p2_dh & p1_dh + p2_dh & p1_dl +
        (1 + p1_dh + p1_dl | p1_name)
  Distribution: Distributions.Bernoulli{Float64}
  Link: GLM.LogitLink()

   logLik    deviance     AIC       AICc        BIC
 -4683.6749  9367.3498  9393.3498  9393.3862  9487.0842

 Number of obs: 10000; levels of grouping factors: 10

Fixed-effects parameters:
-------------------------------------------------------
                    Coef.  Std. Error       z  Pr(>|z|)
-------------------------------------------------------
(Intercept)     2.42733      0.285026    8.52    <1e-16
p1_dh          -3.90603      0.267717  -14.59    <1e-47
p1_dl           1.98044      0.242981    8.15    <1e-15
p2_dh           0.0238098    0.172905    0.14    0.8905
p1_dh & p1_dl  -0.0180805    0.144527   -0.13    0.9004
p2_dh & p1_dh  -0.104624     0.1659     -0.63    0.5283
p2_dh & p1_dl   0.0632909    0.130728    0.48    0.6283
-------------------------------------------------------
```

Table S20: Socially-blind RL Model 3

```
Likelihood-ratio test: 3 models fitted on 10000 observations
---------------------------------------------------------
     DOF  dDOF      LogLik   Deviance   Chisq  p(>Chisq)
---------------------------------------------------------
[1]   10        -4684.0539  9368.1077
[2]   11     1  -4683.9046  9367.8092  0.2985     0.5848
[3]   13     2  -4683.6749  9367.3498  0.4594     0.7948
---------------------------------------------------------
```

Table S21: Socially-blind RL, likelihood ratio tests

Socially-Blind RL, conditioned on winning

```
Generalized Linear Mixed Model fit by maximum likelihood (nAGQ = 1)
  p1_chose_high ~ 1 + p1_dh + p1_dl + p1_dh & p1_dl + (1 + p1_dh + p1_dl | p1_name)
  Distribution: Distributions.Bernoulli{Float64}
  Link: GLM.LogitLink()

   logLik    deviance     AIC       AICc        BIC
 -1020.9555  2041.9110  2061.9110  2061.9589  2126.2730

 Number of obs: 4611; levels of grouping factors: 10

Fixed-effects parameters:
```

```
-------------------------------------------------------
                  Coef.  Std. Error       z  Pr(>|z|)
-------------------------------------------------------
(Intercept)     2.71824    0.37744     7.20    <1e-12
p1_dh          -7.32848    0.40861   -17.94    <1e-71
p1_dl           5.06492    0.488982   10.36    <1e-24
p1_dh & p1_dl   1.3265     0.393813    3.37    0.0008
-------------------------------------------------------
```

Table S22: Socially-blind RL, conditioned on winning 1

```
Generalized Linear Mixed Model fit by maximum likelihood (nAGQ = 1)
  p1_chose_high ~ 1 + p1_dh + p1_dl + p2_dh + p1_dh & p1_dl + (1 + p1_dh + p1_dl | p1_name)
  Distribution: Distributions.Bernoulli{Float64}
  Link: GLM.LogitLink()

   logLik    deviance     AIC       AICc        BIC
 -1009.4648  2018.9297  2040.9297  2040.9871  2111.7279

 Number of obs: 4611; levels of grouping factors: 10

Fixed-effects parameters:
-------------------------------------------------------
                  Coef.  Std. Error       z  Pr(>|z|)
-------------------------------------------------------
(Intercept)     1.98854     0.407244    4.88    <1e-05
p1_dh          -7.57423     0.418223  -18.11    <1e-72
p1_dl           4.86954     0.489243    9.95    <1e-22
p2_dh           0.743444    0.156709    4.74    <1e-05
p1_dh & p1_dl   1.44141     0.395446    3.65    0.0003
-------------------------------------------------------
```

Table S23: Socially-blind RL, conditioned on winning 2

```
Generalized Linear Mixed Model fit by maximum likelihood (nAGQ = 1)
  p1_chose_high ~ 1 + p1_dh + p1_dl + p2_dh + p1_dh & p1_dl + p2_dh & p1_dh + p2_dh & p1_dl +
        (1 + p1_dh + p1_dl | p1_name)
  Distribution: Distributions.Bernoulli{Float64}
  Link: GLM.LogitLink()

   logLik    deviance     AIC       AICc        BIC
 -1001.5546  2003.1091  2029.1091  2029.1883  2112.7797

 Number of obs: 4611; levels of grouping factors: 10

Fixed-effects parameters:
-------------------------------------------------------
                    Coef.  Std. Error       z  Pr(>|z|)
-------------------------------------------------------
(Intercept)      2.21703     0.67793     3.27    0.0011
p1_dh          -10.2556      0.973314  -10.54    <1e-25
p1_dl            8.11524     1.01353     8.01    <1e-14
p2_dh            0.623396    0.453692    1.37    0.1694
p1_dh & p1_dl    1.5711      0.408765    3.84    0.0001
p2_dh & p1_dh    1.84447     0.618233    2.98    0.0029
p2_dh & p1_dl   -2.38949     0.621893   -3.84    0.0001
-------------------------------------------------------
```

Table S24: Socially-blind RL, conditioned on winning 3

```
Likelihood-ratio test: 3 models fitted on 4611 observations
-----------------------------------------------------------
```

```
     DOF  dDOF      LogLik   Deviance    Chisq  p(>Chisq)
-------------------------------------------------------
[1]   10        -1020.9555  2041.9110
[2]   11     1  -1009.4648  2018.9297  22.9813     <1e-05
[3]   13     2  -1001.5546  2003.1091  15.8205     0.0004
-------------------------------------------------------
```

Table S25: Socially-blind, conditioned on winning: Likelihood Ratio Tests