Flexible control of behavioral variability mediated by an internal representation of head direction

Chuntao Dan[1], Brad K. Hulse[1], Vivek Jayaraman[1*], and Ann M. Hermundstad[1*]

[1]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA

*correspondence to: vivek@janelia.hhmi.org, hermundstada@janelia.hhmi.org

ABSTRACT

Internal representations are thought to support the generation of flexible, long-timescale behavioral patterns in both animals and artificial agents. Here, we present a novel conceptual framework for how *Drosophila* use their internal representation of head direction to maintain preferred headings in their surroundings, and how they learn to modify these preferences in the presence of selective thermal reinforcement. To develop the framework, we analyzed flies' behavior in a classical operant visual learning paradigm and found that they use stochastically generated fixations and directed turns to express their heading preferences. Symmetries in the visual scene used in the paradigm allowed us to expose how flies' probabilistic behavior in this setting is tethered to their head direction representation. We describe how flies' ability to quickly adapt their behavior to the rules of their environment may rest on a behavioral policy whose parameters are flexible but whose form is genetically encoded in the structure of their circuits. Many of the mechanisms we outline may also be relevant for rapidly adaptive behavior driven by internal representations in other animals, including mammals.

INTRODUCTION

Behavior often depends on the transformation of sensory information into motor commands based on an animal's internal needs. Some direct responses to sensory stimuli do not require a brain [1] or even neurons [2], but neural networks enable animals to more precisely direct their actions, and to adapt their responses to sensory stimuli based on context, internal state, and experience [3–5]. However, sensory cues are not always reliable or even available, and many animals have evolved the ability to behave still more flexibly by generating and using internal representations of their relationship to their surroundings [6, 7]. These internal representations —for example, those carried by head direction (HD), grid, and place cells [8, 9]— are often tethered to sensory cues, but they allow animals to achieve behavioral goals without depending directly on those cues. Thus, goal-oriented behavior can sometimes operate in two phases: first, a latent learning phase in which an animal explores and acquires an internal model of the structure of its environment, and second, a phase in which the animal modifies its behavior in that environment through its good or bad experiences in specific situations and/or places [10]. Many studies of learned behavior and its neural correlates, particularly those involving mammals, focus on the second phase, using trained animals that have already learned the basic structure of tasks and environments; in doing so, they study task performance more than task acquisition (but see [11–15]). By contrast, in many natural settings, animals must develop internal representations of environmental structure at the same time that they discover the environment's rules through reinforcement of specific actions or contexts. Further, they must use these still-evolving representations to select appropriate actions (or, in the parlance of reinforcement learning (RL) [16], to guide their behavioral policy). Of course, animals in natural settings do not already know the exact tasks that they will be required to perform. Rather, they must often infer the specifics of their behavioral goals based on the consequences of their actions. Moreover, behavioral goals can themselves change over time based on environmental conditions and internal state, requiring animals to balance exploitation of a good situation with explorations away from it.

In this study, we delve into the dynamic process by which internal representations, goals, and behavioral policies develop and work together in the context of rapid, visually-guided operant learning in the fly, *Drosophila melanogaster* [17]. We used a variant of a classical learning paradigm for tethered flies [18] to explore the behavioral policy —and the underlying circuit architecture— that enables flies to modify their actions in response to heat punishment associated with one of two repeating visual patterns arranged symmetrically around the fly. We first quantitatively analyzed the behavior of individual tethered flies to understand the structure of their actions and isolated a set of control parameters that govern the action selection process. We compared how these control parameters should

optimally be structured in order to maintain a behavioral preference for a specific visual pattern, which is repeated in this environment, versus a preference for a specific heading. We showed that flies' behavior is consistent with a heading-dependent policy, but only if the symmetry of the setting evokes a predictable instability in the internal representation of heading. Recent studies have shown that the fly's internal representation of head direction (HD) is observable as a 'bump' of calcium activity in so-called compass or EPG neurons in the ellipsoid body (EB) [19], a substructure of an insect brain region called the central complex (CX) [20–26]. We used two-photon calcium imaging of compass neurons in tethered flying flies to examine the HD representation in the visual scene used in our behavioral paradigm. We found that the symmetry of the visual setting indeed induces a structured instability in the fly's HD representation. We used these observations, together with existing physiological, behavioral, and anatomical observations [19, 27–34], to construct a model for how CX circuits downstream of compass neurons might use the HD representation to both learn a goal heading in its visual environment and select actions that are driven by the fly's current heading relative to its goal. Importantly, the structure of the fly's behavioral policy relative to this goal heading appears to be hardwired rather than developed from scratch, allowing rapid learning and immediate adjustments to new goal headings. Finally, we used the model to understand the interplay of the processes governing the generation of the HD representation in this environment, the inference of the goal heading, and the selection of actions. Our results and conceptual framework cast decades of influential fly visual learning studies [18, 27, 35, 36] in a new light, suggesting that learning in these tasks operates on an internal HD representation rather than directly associating actions with reinforced visual patterns. These results also provide a window into how rapid behavioral flexibility can be enabled by a prestructured yet adaptive policy tethered to an internal representation, and they underscore the importance of the reliability of that representation in shaping the animal's behavior.

RESULTS

Tethered flying flies change their visually-guided behavior after thermal conditioning

To explore whether and how flies adapt their behavior in response to aversive conditioning, we modified a classical visual learning paradigm that was developed in the Heisenberg lab several decades ago [18]. In our modified paradigm, tethered flying flies in an LED arena were given closed-loop control of their angular orientation relative to a visual scene by locking angular rotations of visual patterns on the arena to differences in flies' left and right wingbeat amplitude, a proxy for their intended yaw movements (Fig 1a, upper panel) [37, 38] (see Methods). We used a periodic visual scene consisting of four quadrants of horizontal bars (Fig 1a, middle panel). In two opposing quadrants, bars were positioned at a low elevation; in the other two quadrants, bars were positioned at a high elevation. We assessed flies' naive preferences for different quadrants of the visual scene during a pair of 2-min-long 'naive trials' (Fig 1a, lower panel). During subsequent 'training trials', two symmetric quadrants (the 'danger zone') of the visual scene were paired with an aversive heat punishment delivered via an infrared laser to the abdomen of the fly; the remaining two quadrants (the 'safe zone') were left unpunished. In 'probe trials' with no heat punishment, we assessed whether flies formed lasting associations between different quadrants of the arena and the aversive heat.

Prior to training, flies spent a similar amount of time in different parts of the arena (Fig 1b, upper row). Over the course of training, the residency increased within the safe zone and decreased within the danger zone (Fig 1b, left column); this was not observed in control flies that did not receive laser punishment ('no-laser' controls; Fig 1b, right column). We quantified performance via a metric used in classical studies, the performance index ('PI score') [35, 39], which measures the relative fraction of time spent in safe versus dangerous quadrants; larger PI scores indicate a stronger preference for safety. We found, consistent with classical studies [27, 35, 39], that laser-trained flies, on average, learn within minutes to avoid quadrants associated with punishment; this was reflected in an increase in PI scores during training trials that was maintained in probe trials (Fig 1c, shaded bars). No-laser control flies did not significantly change their preference for either quadrant (Fig 1c, white bars).

A probabilistic policy captures tethered flies' visually-guided behavior

What precise changes in behavior underlie this operant learning [17, 40]? To answer this, we sought to construct a generative model of behavior, or behavioral policy [16], that could account for naive and conditioned behavior. We noted that PI scores do not capture the different types of behavioral trajectories that flies generate in this paradigm,

with or without conditioning (Fig 2a,b). Thus, our goal was to use a behavioral policy to characterize flies' behavioral variability in the absence of conditioning and to then predict how this variability should change based on experience.

Flies' behavior can be quantified in terms of the execution of different modes of patterned movement [41, 42]. During free and tethered flight, flies exhibit periods of fixation during which they maintain a near-constant heading over time [37, 41, 42]. These fixations are often punctuated by body 'saccades', or ballistic turns, that result in abrupt changes in heading [41, 42]. We observed these behavioral modes in both laser-trained (Fig 2a) and no-laser control (Fig 2b) flies (note that there were no obvious differences in the distribution of fixation durations between laser-trained and no-laser control flies; see SI Fig S1). We approximated behavior as being composed of only these two modes (Fig 2c,d), and we used behavioral kinematics to determine transitions between them (SI Fig S2a; see Methods). Individual fixations tended to be long in duration with near-zero average angular velocity, while individual saccades were distinguished by high angular velocity over short durations (SI Fig S2b). We used the variability in these properties across flies and trials to infer a generative model of behavior in which flies control the distribution of possible actions within each of these two modes. Specifically, our analysis supports a generative model in which flies control the relative probability of initiating clockwise versus counterclockwise saccades through an adaptive rotational bias, and the average duration of fixations through a drift-diffusion process with an adaptive drift rate (Fig 2e, SI Fig S2c-j; see Methods for a detailed analysis of additional control parameters).

We then asked how these two adaptive control parameters —rotational bias and drift rate— should be tuned as a function of the fly's orientation in the arena. Flies, like many other insects, display individual heading preferences, maintaining a specific 'goal heading' for periods of time (SI Fig S3) [28, 31, 43], a behavior that is thought to aid dispersal and long-range navigation [44–49]. However, rather than purely fixating on one goal heading, both walking and flying flies also explore other headings while centering their explorations around the goal heading [28, 31, 43, 50–52], a behavioral pattern that matches our observations in this paradigm. These results led us to compare two alternative hypotheses: one in which flies' actions were linked directly to specific visual patterns, as has been suggested to explain fly behavior in this paradigm (see, for example, [53]), and one in which flies' actions were linked to an internal goal heading. When model flies were trained through an RL algorithm to maintain a preference for a specific visual pattern (Fig 2f), they learned to fixate longer when orienting towards the visual pattern, and they biased the direction of their saccades towards the visual pattern, regardless of its specific orientation within the arena. In a scene with two sets of repeating visual patterns, this resulted in a bimodality in both the duration of fixations and the directionality of saccades (Fig 2f). In contrast, when model flies were trained to maintain a goal heading, their fixations and saccades were structured with respect to a single orientation within the arena (Fig 2g). When we aligned the behavioral data to individual flies' preferred heading in the arena ('arena heading') (Methods), we observed that the behavior exhibited a bimodal structure, with flies locally directing their turns towards, and fixating longer at, the two headings that correspond to symmetric views of the visual scene (Fig 2h). However, rather than exhibiting the symmetric structure that we would have expected for a visual-pattern-based policy (Fig 2f), the structure in the behavior was stronger at the location of the preferred heading compared to the $180°$-symmetric location (Fig 2h). That is, there were asymmetries in the duration of fixations and in the probabilistic bias of saccades at two visually indistinguishable headings. Thus, flies' behavioral patterns seemed not to qualitatively fit either of our hypotheses.

Instability in flies' internal heading representation impacts its behavioral policy

Flies' flexibility in heading preferences depends on compass neurons [28, 31, 43], and inputs to the compass neurons have been linked to flies' ability to remember specific orientations relative to visual patterns [36, 54]. Further, CX circuits downstream of compass neurons have been implicated in a behavioral paradigm similar to ours [27]. However, the compass neuron HD representation —equivalent to the fly's 'internal heading' in this setting (see Supplemental Information for further discussion of this issue)— is unstable in visual environments with two prominent features placed $180°$ apart [19, 29, 30]. We therefore hypothesized that the observed asymmetry in flies' behavior might arise from such an instability in this visual setting as well. Indeed, when we coupled the goal-heading-dependent policy to an unstable heading representation (Fig 2i; see Methods and SI Fig S4), the resulting behavioral readout exhibited an asymmetric bimodal structure (Fig 2i, blue box, right column) that mimics the actual behavioral data (Fig 2h). The fact that we observed this structure in both laser-trained and no-laser control flies (Fig 2h, left and right columns, respectively) suggests that their behavior is governed by a "prestructured" internal policy tethered to the difference between flies' current internal heading and a single internal goal heading, and that even untrained flies have a goal heading [28, 31, 43]. We hypothesized that this unimodal policy only

manifests in a bimodal structure due to the impact of the symmetric nature of the visual environment on the fly's internal heading representation.

Does the CX's heading representation exhibit a structured instability in the visual setting used in our behavioral paradigm as we hypothesize? To address this question, we monitored the compass neuron heading representation in the EB (Fig 3a) using two-photon calcium imaging in tethered flies flying in a visual setting similar to that used in the learning assay (Fig 3b; see Methods for details). In any given visual environment, the heading representation tethers to the visual scene with a particular 'offset' that defines the orientation of the heading bump relative to the orientation of the scene [19, 29, 30] (schematized in Fig 3c,d). This offset is arbitrary, varies from fly to fly, and can vary over time as well [19, 28–31]. In a previous study, we showed that the stability of the mapping between a visual scene and the heading representation depends on particular characteristics of that scene; briefly, scenes whose rotational auto-correlation produce single, dominant peaks tend to induce stable, one-to-one mappings from the visual scene onto the heading representation [29]. In contrast, the visual scene employed in our behavioral experiments is two-fold symmetric, with two peaks in its rotational auto-correlation. We predicted that this scene, which matches those commonly used in classical visual learning experiments [18, 27], would induce an instability in the offset between the heading bump and the visual scene, as has been observed in visual scenes with two identical vertical stripes placed at opposing orientations [19, 29, 30, 55].

When we tested the stability of the offset of the heading representation relative to the visual scene, we found, as predicted, that the heading bump tended to "jump" between two offsets that reflected symmetric views of the scene (Fig 3e,f). In our two-fold symmetric scene, this corresponds to a jump of 180° (Fig 3g). Examining offsets across flies (Fig 3h), we found that the distribution of offsets was bimodal in a majority of flies (Fig 3i, top panel), with peaks separated by 180° (Fig 3i, bottom panel; see Supplemental Information for discussion of the smaller peak at 90°). In correspondence with this, we found that different wedges of the EB were active at symmetric angular orientations, and thus their heading tuning curves had two peaks also separated by 180° (Fig 3j). This resulted in a two-to-one mapping from the visual scene onto the heading representation, similar to the tuning previously observed in simpler symmetric scenes [19, 29, 30, 55]. We next asked whether bump jumps were more likely to occur at certain locations in the EB (Fig 3k). We measured how frequently the bump tended to jump from different locations in the EB, relative to the number of visits the bump made to that location. When we analyzed the location of these jumps relative to the location of maximal residency in the EB, we found that jumps were least likely to occur at this preferred bump heading, and most likely to occur 180° away from this preferred bump heading (Fig 3l). Together, this suggests that the symmetries of the scene induce instabilities [19, 29, 30, 55] that manifest in jumps of the heading bump between locations in the EB that correspond to symmetric views of the visual scene. In a scene with two-fold symmetry, this instability is not uniform around the EB, but is strongest at the location symmetric to the preferred bump heading (Fig 3l), as predicted (Fig 2i). Taken together, these results suggest that the observed structure in flies' behavior (Fig 2h) is not a direct result of a behavioral policy tethered to specific visual patterns. Rather, it is a result of the impact of the scene's symmetry on the dynamics and stability of the heading representation, and of a behavioral policy tethered to that representation.

How might flies' performance be affected if reinforcement is coupled to these same symmetries? Under simplifying assumptions, the heading instability would not hinder performance so long as the most stable bump heading is aligned with the goal heading (SI Fig S5). In such a setting —where both the heading representation and the behavioral policy remain structured with respect to a single goal heading— learning need only shift the location of the goal heading to enable the fly to quickly adapt to new environments.

A simple circuit model implements a prestructured goal-driven behavioral policy

To understand how the flies' circuitry could ensure that both the heading instability and behavioral policy remain structured with respect to the goal heading (Fig 4), and how learning acts to shift these structures (Fig 5), we constructed an abstract circuit model that expands upon the algorithm shown in Fig 2h. This model combines insights from existing models of the heading system [29, 56–60] with conceptual insights from the CX connectome [34] (Fig 4a, upper panel), but it intentionally abstracts away much of the known detail of CX circuit structure and function to focus on key computations that we believe underlie flies' behavior in this assay (see SI Fig S6 and Supplemental Information for a detailed description of how various features of our model relate to existing anatomical and functional data). Briefly, our model uses five neuron types that combine activity related to the fly's current heading and goal heading in order to drive behavior (Fig 4a, lower panel). Compass neurons maintain a single, sinusoidal bump of activity that tracks the fly's current heading; this sinusoidal profile and the bump's

responsiveness to the fly's turns are implicitly assumed to be maintained by a ring attractor network and by neuron types that we did not explicitly model here [29, 34, 43, 58, 61–63]. The stability of this heading representation is determined by a set of inhibitory "compass weights" that captures the relationship between the visual scene and the bump [29, 30, 60, 64]. A set of goal neurons maintains activity related to the fly's goal heading; this activity, which can take on an arbitrary profile, is read out directly from a set of excitatory "goal weights". The heading and goal activity is combined in three downstream populations of action neurons whose net output is used to control the duration of fixations and the direction of saccades.

The prestructured behavioral policy is enforced by the three populations of action neurons. These action neuron populations each inherit the heading bump from their upstream inputs, but, from the perspective of subsequent computations, with a phase shift relative to the compass neurons (Fig 4b). That is, considering the sinusoidal heading bump as a phasor, which captures both the bump's amplitude and its angular orientation within the 360° span of the EB, the different action neuron populations each have bumps with specific angular differences relative to the compass neuron bump (see Supplemental Information for the anatomical basis of this assumption). The action neurons compute the net overlap between their different phase-shifted versions of heading activity and the goal activity (Fig 4b). Individual neurons in each population multiplicatively combine their inputs, and the summed output activity of each population is used to drive two downstream controllers that initiate CW or CCW rotations (Fig 4c). Two of these populations receive versions of the current heading bump that are phase-shifted by -90° and $+90^\circ$, and they project unilaterally to the CW and CCW controller, respectively (see Supplemental Information for the relevant CX neuron types); these neurons thus control the initiation of turns [32, 34, 59]. The third population receives a version of the current heading bump that is phase-shifted by 180° [34] and projects bilaterally to both controllers, thereby controlling fixation. The summed output of the multiplicative operation between a sinusoidal bump profile and an arbitrary goal profile guarantees that the output of each action population varies sinusoidally as a function of the angular difference between the current heading bump and the circular mean of the goal activity, regardless of the specific profile of goal activity; thus, the circular mean of the goal activity specifies the goal heading (see Eq. 30 for a brief mathematical explanation). As a result, the two populations of action neurons that control turns will have the largest output —and thus would most likely drive CW or CCW saccades— when the fly's current heading bump is 90° to the right or the left of the goal heading, respectively (note that the bump is tethered to the movement of the visual scene, and moves opposite to the direction of the fly's turn). The fixation population will have the largest output —and thus would most likely drive short fixations with a high drift rate— when the fly's current and goal headings are anti-aligned. Together, this architecture ensures that the fly's behavior remains structured with respect to the angular difference between the current and goal headings, regardless of their absolute orientations. The shape and range of the goal activity determines how strongly the goal heading drives fixational and saccadic behavior; the stronger the circular mean of the goal activity profile, the larger the difference in behavior at the goal and anti-goal headings, and thus the more structured the behavioral output (SI Fig S7). Within the constraints of the learning algorithm considered here, a high degree of behavioral structure is achieved by driving the goal weights towards a sinusoidal profile (see Eq. 30 and the following discussion for a brief mathematical explanation).

The compass weights that define bump instability at different orientations (Fig 4d) represent the summed impact of inputs to compass neurons from visual "ring neurons" (Fig 4e) [65–67]. Visual ring neurons have feature-tuned receptive fields that tile space (Fig 4e, upper panel; [65]), and they synapse onto compass neurons via all-to-all inhibitory connections in the EB [34] (Fig 4e, lower panel; also see Supplemental Information). During exploration of a visual scene, inhibitory Hebbian-like plasticity is thought to weaken synapses from active ring neurons onto active compass neurons at the location of the heading bump in the EB, thus creating a consistent mapping between the visual scene and the heading representation (heatmap in Fig 4e; [29, 30]). We approximate this full weight matrix by a single vector of compass weights that captures the summed (net) inhibition from the population of ring neurons onto each compass neuron. The heading bump will prefer to occupy regions of the EB that are only weakly, rather than strongly, inhibited by active ring neurons, and thus this set of weights determines the stability of the heading bump [29]. In a scene with two-fold symmetry, the same populations of ring neurons will be active at two orientations of the scene that we depict as corresponding to two potentially active compass neurons in the EB (Fig 4f). If this set of active ring neurons produces the same net inhibition onto these two compass neurons, the heading bump will exhibit a bistability in which it could occupy either of these two EB locations with equal probability (Fig 4f; left column). However, if the same set of active ring neurons produces *different* levels of net inhibition onto these two compass neurons, the heading bump will prefer the location with weaker net inhibition, and will be more likely to preferentially jump towards this location (Fig 4f; right column). For a two-fold symmetric visual scene, the difference in net inhibition between two locations in the EB separated by 180° will determine the probability that the heading bump will jump between these headings. As a result, different compass weight profiles, capturing different

distributions of net inhibition across the compass neuron population, will lead to different patterns of instability around the EB (SI Fig S7); the shape of the weight profile controls how likely the bump is to jump from each orientation and how likely it is to jump towards, rather than away from, the most stable bump heading. However, it is important to note that the compass weight profile does not, in general, determine whether or not a specific pair (or more) of compass neurons might compete for the bump. The presence of multiple similar visual features in a given scene evokes similar ring neuron activation patterns at multiple headings; this similarity in ring neuron activation patterns then triggers a competition in multiple compass neurons that are tuned to those headings. In the symmetric visual setting used in our behavioral paradigm, these different headings happen to be separated by $180°$.

Circuit model with heading-dependent policy captures experimental observations

The compass weight profile is shaped by the residency of the bump (likely mediated by motor-state-dependent neuromodulatory input that we do not explicitly model, see Supplemental Information and Methods); the longer the bump resides in a particular location at a given scene orientation, the lower the local compass weight profile will be, and the higher the probability that the bump will jump to this location in the future. During exploration of a new visual scene, the structure of the behavioral policy controls the fly's heading, and, therefore, also the likelihood of the compass bump favoring some EB locations over others. In our model, we assumed that the summed inhibitory compass weights are weakened at the location of the heading bump via an inhibitory Hebbian-like plasticity rule (exemplified in the top panel of Fig 4g; Methods). A strong goal heading will drive a prestructured behavioral policy, which in turn will drive the heading bump to most frequently occupy EB locations near the goal heading (Fig 4g, middle panels). Over time, the heading instability will develop a sinusoidal structure that is tethered to the policy, such that the most stable bump heading is aligned with the goal heading (Fig 4g, bottom panel).

Together, these key features —a prestructured behavioral policy and a structured bump heading instability aligned to the policy— reproduce the signatures of heading bump dynamics and behavior shown in Figs 2-3. In simulated model flies with fixed heading and goal weights (Fig 4h), the heading bump jumps least frequently at the goal heading (Fig 4h, lower left), similar to what we observed in Fig 3l. This, in turn, produces the expected bimodality in compass neuron tuning (Fig 4h, lower right), similar to what we observed in Fig 3j. When this unstable heading is used to drive behavior via the fly's internal policy, the behavioral readout exhibits the expected bimodality (Fig 4h, middle right), again similar to what we observed in real flies (Fig 2h).

Shifts in goal headings during learning are predicted by the circuit model

In our visual learning paradigm, a new goal heading must form alongside a new heading representation. We assume that the excitatory goal weights are updated based on heading input and on valence input by neuromodulatory neurons that strengthen or weaken synapses in proportion to the activity of the heading bump when the fly is orienting towards positively or negatively reinforced headings, respectively (Fig 5a; see Supplemental Information for candidate neuromodulatory neuron types that could serve such a role [34, 68]). When a single model fly begins with weak and unstructured goal weights, it evenly samples different headings. Through training, both the goal heading and the most stable bump heading begin to stabilize at the same location (Fig 5b-c). As the range of the goal weights increases, the motor drive becomes more structured (dark purple and green curves in right column of Fig 5b), which manifests in a more structured exploration of arena and bump headings (Fig 5d, middle and lower right panels). This sharpens the compass weight profile, in turn driving the heading bump more strongly to a single goal heading (Fig 5d, middle and lower left panels). We observe qualitatively similar behavior across many different initializations of compass and goal weights (Fig 5c); over time, both the range and location of compass and goal weights become correlated with one another, and eventually lock to one of the two safe zones. On average, model flies increase the range of their goal weights and shift their goal heading towards safety, which results in an increase in their PI scores (Fig 5e). Across model flies, the initial range of goal weights (which controls the initial degree of structure in the behavior) more strongly impacts changes in performance than does an initial heading preference for safety versus danger; model flies that began with a larger range of goal weights (and thus with more structure in their behavior) showed larger changes in PI scores, whereas initial heading preferences only weakly impacted changes in performance (Fig 5f; see SI Fig S8 for a more detailed explanation of how initial range and preference impact changes in performance).

These qualitative features of the model are consistent with the behavior of single flies that exhibited large changes in performance after training (Fig 5g-h). Fig 5g illustrates changes in residency over the course of training for two

flies that began with either a weak preference (left column) or a strong preference for the edge of the danger zone (right column). Flies' changes in residency are accompanied by an increase in the structure of the behavior with respect to the transient goal heading (Fig 5h, upper left), a shift in the location of the goal heading (Fig 5h, middle left), and an overall increase in PI scores (Fig 5h, lower left), all of which are observed in group-averaged data (Fig 5h, right column) and predicted by the model (Fig 5e). Flies that began with more structured behavior in naive trials showed larger average changes in PI scores during later trials (Fig 5i, upper panel), whereas initial preference did not strongly impact average changes in PI scores (Fig 5i, lower panel), again consistent with model predictions (Fig 5f). Together, these results suggest that flies' performance depends on the degree of initial structure in their behavior, and evolves over time as flies simultaneously map their sensory surroundings and their goals within them.

DISCUSSION

We analyzed *Drosophila* behavior in a modified version of a classical visual learning paradigm [17, 18], and combined this analysis with calcium imaging of compass neurons in a similar visual setting to develop an abstract circuit model of the computations underlying this operant behavior. Our model enabled us to examine how flies use evolving heading representations to guide their behavior towards goals that they simultaneously infer from heading-specific heat reinforcement. Based on previous physiological, behavioral, connectomic, and modeling studies focused on the CX [19, 28–32, 34, 55–57, 59–61], and a combination of the imaging and behavioral results in this study, we suggest that flies rely on a prestructured behavioral policy tethered to a flexible internal heading representation that specifies controllable properties of the fly's actions relative to a single goal heading. Flies operantly learn to shift this goal heading based on reinforcement, quickly redirecting their actions toward unpunished headings. However, because the heading representation on which this policy is built is learned alongside the goal heading, any inaccuracies in this representation impact the fly's behavior, which, in turn, impacts the fly's ability to sample the new environment and appropriately update its goal heading. We used artificial visual environments with repeating patterns tethered to aversive punishment to reveal this interdependence of representational stability, goal learning, and goal-directed behavior. Our framework makes clear how symmetries in the visual scene induce the fly's heading bump to jump, and how this instability, when appropriately structured, does not interfere with the fly's ability to learn in this visual setting. Because we study the learning process beginning with the fly's first encounter with this sensory environment, the formation of the heading representation is intrinsically coupled with the formation and shifting of the goal heading. As a result, instabilities in the heading representation can impact learning, something that we see play out at the level of flies' changes in performance.

Our results may warrant a reinterpretation of decades of studies in visual pattern, spatial, and color learning in tethered flying flies [18, 27, 35, 40]. In all these studies, flies were believed to have learned to associate their actions with specific visual features or objects; these conclusions were based on the near-symmetric structure of fly behavior in settings with the same patterns in opposing quadrants of circular arenas. This symmetry of the visual environment was believed to rule out the possibility of learning based on heading. We suggest instead that flies build heading representations tethered to these different visual surroundings, and rely on heading-representation-based goal learning to associate rewards or punishments with different headings. We further argue that any subtle asymmetries in flies' responses to visual patterns were caused by an instability in the heading representation that arose from a symmetry of the scene. This instability rendered irrelevant any limitations that a single-goal-heading-based policy might otherwise impose on the fly's actions in a setting with multiple safe and dangerous headings. It is possible that such heading instabilities would be less frequent in free flight, where proprioceptive cues are likely to play a greater role in controlling heading bump dynamics (see, for example, [55]). Whether flies can learn multiple distinct goal headings, and to what extent they can form more complex policies beyond the prestructured single-goal-heading policy invoked here, is not yet known. Performance in a place-learning paradigm suggests that flies can learn more complex associations to guide navigation through 2D environments [69]. It is also possible that, in contrast to the CX-based learning we study here, many spatial navigation behaviors may rely on associations made in the mushroom body [23, 70–76], a brain region that has also been suggested to be involved in some operant visual learning tasks in tethered flying flies [35, 77].

For most of our analyses, we decomposed the tethered fly's behavior into two different modes: fixations and saccades. Freely flying flies are known to exhibit these different modes [78] that are characterized by distinct kinematic properties and necessitate both continuous and discrete control [42, 79]. The same modes have also been observed previously in tethered flying flies [42]. We explicitly incorporated these modes into the construction of a behavioral policy, and we used the observed variability in kinematic parameters across flies and trials to infer the parametric form and control parameters of this policy. When combined with optimal RL algorithms [16], this

enabled us to specify how these control parameters should change based on experience. This approach bears similarities to recent studies proposing that learning operates on generative parameters that control distributions of movements, rather than on the higher-dimensional space of all possible movements [80–82]. We then used this approach to specify how these control parameters should be structured as a function of the fly's current heading to maintain a goal heading over time. Rather than learning this structured relationship from scratch, we showed that this relationship is pre-built into how untrained flies sample their surroundings, a strategy that might facilitate dispersal in the absence of explicit goals [50]. Indeed, these same patterns of structured behavior resemble those observed in tethered behaving flies responding to visual features that are innately attractive or aversive [79]; note that an additional parameter, the size of saccades, also varies based on angular distance from a "goal" object in those visual environments, something that was less striking our visual setting. We suggest that flies may, in fact, rely on different visuomotor pathways in different settings. Responses to innately attractive or aversive objects [83–87] could rely on direct and hardwired visuomotor pathways that recruit banks of feature detectors in the optic foci [88, 89] and, perhaps, relatively stereotyped motor responses dependent on the spatial receptive fields of feature detector inputs [79]. We suggest that learned responses, in contrast, rely on an indirect pathway that recruits prestructured probabilistic behavioral policies tethered to the relationship between internal representations of current and goal headings. We did not explore whether these probabilistic biases could be strengthened through longer training protocols. We note that although it might seem optimal to steer towards and then maintain a single goal heading rather than initiate directed turns that are probabilistically biased toward this heading, using such a default behavioral strategy would likely be too predictable to avoid predation [90] and would minimize exploration. Indeed, many animals, including flies, show stochasticity in their actions when behaving freely in dynamic settings [90, 91]. In the fly, recent evidence from the connectome suggests that the architecture of columnar neurons in the FB could implement a prestructured behavioral policy for steering towards a goal heading by using the fly's current bump heading [34, 59] (see also Supplemental Information). Importantly, within this model, learning acts to modify the location and strength of the goal heading while preserving the entire structured relationship between different control parameters across different headings. Our results suggest that these associations are mediated via a flexible pathway through the CX; however, direct sensorimotor pathways that instruct reflexive actions might, in fact, work in concert with these flexible pathways, and in the aversive conditioning setting considered here, might enable the fly to quickly escape punished zones. How such pathways are balanced to guide reflexive and flexible actions, and whether the outcome of reflexive actions can be used to inform future flexible actions, is not yet known (but see Supplemental Information for a discussion of how reflexive actions might be incorporated in the framework presented here).

In contrast to many behavioral paradigms in mammals, flies in this paradigm learn within a matter of minutes —without shaping or instruction— to direct their behavior away from punishment and towards safety. Our results suggest that flies' rapid learning relies on a strong inductive bias in the form of a prestructured behavioral policy that dictates flies' sampling strategy. This prestructured policy effectively assumes the existence of a single goal heading for the fly, and efficiently directs them towards identifying and orienting towards such a heading. Such inductive biases reduce the possible scenarios that are explored during learning and can thereby speed up the learning process when these scenarios are compatible with the learning task [92, 93]. Recent RL studies have explored how such inductive biases might be constructed by learning common features across different learning tasks [94, 95], a process known as learning to learn [96]. Here, we show how an inductive bias that is likely learned over evolutionary timescales can be inferred directly from an animal's behavior in the absence of an explicit task. The ability to rapidly exploit this inductive bias, in this case by shifting a single goal heading, relies on faster-timescale learning. Whether, and to what extent, this faster-timescale learning could modify the flies' prestructured behavioral policy —for example, by suppressing the exploratory component through increased training— remains unknown, as does the potential for behavioral state information —for example, walking instead of flight— to switch which actions are controlled through the same behavioral policy.

It has recently been suggested that rapid learning in both artificial and biological systems relies on combining context-dependent memories with efficient exploitation of environmental and task structure [93]. Here, we provide insights into how specific neural circuits might instantiate a behavioral policy that has evolved to address ecological needs through efficient actions, and how this policy both informs and is shaped by a flexible, and perhaps context-dependent, internal representation of the environment. Targeted genetic access to the specific cell types that might mediate this learning provides an avenue for rigorously testing these ideas in the future.

ACKNOWLEDGEMENTS

We thank Jason Wittenbach and Daniel Barabasi for pilot modeling efforts, Parvez Ahammad for useful early discussions, and Alice Wang for pilot experiments. We are grateful to Bjorn Brembs for informative email discussions on the design and interpretation of classical visual learning studies. Eyal Gruntman, Michael Reiser, Josh Dudman, John Tuthill, TJ Florence, Sung Soo Kim, Yoshi Aso, and members of the Jayaraman and Hermundstad labs provided insightful input at different points of the study. We received useful feedback on the manuscript from Sandro Romani, Marcella Noorman, Hannah Haberkern, and Dan Turner-Evans. Ramya Kappagantula (Janelia Project Technical Resources, PTR) contributed a large body of experimental data that was not included here, but will be part of future versions of this manuscript that she will then co-author! We thank Gudrun Ihrke for her expert management of PTR. We thank Dan Milkie (now at Janelia) and Andy Chiu from Coleman Technologies for help with developing the FPGA Wingbeat Analyzer. We thank Janelia Experimental Technology (jET) for technical assistance, especially: Jinyang Liu for the LED arena controller code, Steve Sawtelle for the D2A converter connected to the FPGA Wingbeat Analyzer, and Tanya Tabachnik, Igor Negrashov, and Bill Biddle for designing and manufacturing the fly mounting assembly used for two-photon imaging. We are grateful to Janelia's Drosophila Resources team for stock building and maintenance, and to the Media Prep Facility for special fly food that kept our finicky flies flourishing.

This work was funded by the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

CD and VJ conceived of the study waay back when. AMH, CD, and VJ then reconceptualized the study. CD performed all experiments, data processing, and initial data analysis. AMH performed all behavioral analysis in this manuscript, with input from CD, VJ, and BKH. VJ and AMH analyzed imaging data with input from CD. All authors interpreted results. AMH developed the theoretical framework and performed all modeling and simulations, with conceptual input from VJ, CD, and BKH. The proposed CX circuit implementation of the model was conceived over multiple feedback loops involving VJ, AMH, CD, and BKH, with BKH contributing, in particular, to the CX implementation of the behavioral policy. AMH and VJ wrote the paper, with input and editing from CD and BKH.

COMPETING INTERESTS

We do not have any competing interests.

References

1. Dupre, C. & Yuste, R. Non-overlapping Neural Networks in Hydra vulgaris. *Current Biology* **27,** 1085–1097 (2017).
2. Wadhams, G. H. & Armitage, J. P. Making sense of it all: bacterial chemotaxis. *Nature Reviews Molecular Cell Biology* **5,** 1024–1037 (2004).
3. Huston, S. J. & Jayaraman, V. Studying sensorimotor integration in insects. *Current Opinion in Neurobiology* **21.** Sensory and motor systems, 527–534 (2011).
4. Calhoun, A. J. & Murthy, M. Quantifying behavior to solve sensorimotor transformations: advances from worms and flies. *Current Opinion in Neurobiology* **46.** Computational Neuroscience, 90–98 (2017).
5. Crochet, S., Lee, S.-H. & Petersen, C. C. Neural Circuits for Goal-Directed Sensorimotor Transformations. *Trends in Neurosciences* **42,** 66–77 (2019).
6. Pouget, A & Snyder, L. Computational approaches to sensorimotor transformations. *Nat Neurosci* **3 Suppl,** 1192–1198 (2000).
7. Wolpert, D. M. & Flanagan, J. R. Computations underlying sensorimotor learning. *Current Opinion in Neurobiology* **37.** Neurobiology of cognitive behavior, 7–11 (2016).
8. Knierim, J. J. & Zhang, K. Attractor Dynamics of Spatially Correlated Neural Activity in the Limbic System. *Annual Review of Neuroscience* **35.** PMID: 22462545, 267–285 (2012).
9. Finkelstein, A., Las, L. & Ulanovsky, N. 3-D Maps and Compasses in the Brain. *Annual Review of Neuroscience* **39.** PMID: 27442069, 171–196 (2016).
10. Tolman, E. C. & Honzik, C. H. Introduction and removal of reward, and maze performance in rats. *University of California Publications in Psychology* **4,** 257–275 (1930).
11. Huber, D. *et al.* Multiple dynamic representations in the motor cortex during sensorimotor learning. *Nature* **484,** 473–478 (2012).

12. Poort, J. *et al.* Learning Enhances Sensory and Multiple Non-sensory Representations in Primary Visual Cortex. *Neuron* **86,** 1478–1490 (2015).
13. Peters, A. J., Lee, J., Hedrick, N. G., O'Neil, K. & Komiyama, T. Reorganization of corticospinal output during motor learning. *Nat Neurosci* **20,** 1133–1141 (2017).
14. Coddington, L. T. & Dudman, J. T. Learning from Action: Reconsidering Movement Signaling in Midbrain Dopamine Neuron Activity. *Neuron* **104** (2019).
15. Kuchibhotla, K. V. *et al.* Dissociating task acquisition from expression during learning reveals latent knowledge. *Nat Commun* **10,** 2151 (2019).
16. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* Second (The MIT Press, 2018).
17. Brembs, B. & Heisenberg, M. The operant and the classical in conditioned orientation of Drosophila melanogaster at the flight simulator. *Learn Mem* **7,** 104–115 (2000).
18. Wolf, R. & Heisenberg, M. Basic organization of operant behavior as revealed in Drosophila flight orientation. *Journal of Comparative Physiology A* **169,** 699–705 (1991).
19. Seelig, J. D. & Jayaraman, V. Neural dynamics for landmark orientation and angular path integration. *Nature* **521,** 186–191 (2015).
20. Strauss, R. The central complex and the genetic dissection of locomotor behaviour. *Current Opinion in Neurobiology* **12,** 633–638 (2002).
21. Pfeiffer, K. & Homberg, U. Organization and functional roles of the central complex in the insect brain. *Annu Rev Entomol* **59,** 165–184 (2014).
22. Turner-Evans, D. B. & Jayaraman, V. The insect central complex. *Curr Biol* **26,** R453–457 (2016).
23. Webb, B. & Wystrach, A. Neural mechanisms of insect navigation. *Curr Opin Insect Sci* **15,** 27–39 (2016).
24. Varga, A. G., Kathman, N. D., Martin, J. P., Guo, P. & Ritzmann, R. E. Spatial Navigation and the Central Complex: Sensory Acquisition, Orientation, and Motor Control. *Front Behav Neurosci* **11,** 4 (2017).
25. Heinze, S. Unraveling the neural basis of insect navigation. *Curr Opin Insect Sci* **24,** 58–67 (2017).
26. Honkanen, A., Adden, A., da Silva Freitas, J. & Heinze, S. The insect central complex and the neural basis of navigational strategies. *J Exp Biol* **222** (2019).
27. Liu, G. *et al.* Distinct memory traces for two visual features in the Drosophila brain. *Nature* **439,** 551–556 (2006).
28. Giraldo, Y. *et al.* Sun Navigation Requires Compass Neurons in Drosophila. *Curr Biol* **28,** 2845–2852 (17 2018).
29. Kim, S. S., Hermundstad, A. M., Romani, S., Abbott, L. F. & Jayaraman, V. Generation of stable heading representations in diverse visual scenes. *Nature* **576,** 126–131 (2019).
30. Fisher, Y. E., Lu, J., D'Alessandro, I. & Wilson, R. I. Sensorimotor experience remaps visual input to a heading-direction network. *Nature* **576,** 121–125 (2019).
31. Green, J., Vijayan, V., Mussells Pires, P., Adachi, A. & Maimon, G. A neural heading estimate is compared with an internal goal to guide oriented navigation. *Nat Neurosci* **22,** 1460–1468 (2019).
32. Rayshubskiy, A. *et al.* Neural circuit mechanisms for steering control in walking Drosophila. *bioRxiv* (2020).
33. Shiozaki, H. M., Ohta, K. & Kazama, H. A Multi-regional Network Encoding Heading and Steering Maneuvers in Drosophila. *Neuron* **106,** 126–141.e5 (2020).
34. Hulse, B. K. *et al.* A connectome of the Drosophila central complex reveals network motifs suitable for flexible navigation and context-dependent action selection. *bioRxiv* (2020).
35. Tang, S. & Guo, A. Choice Behavior of Drosophila Facing Contradictory Visual Cues. *Science* **294,** 1543–1547 (2001).
36. Guo, C. *et al.* A conditioned visual orientation requires the ellipsoid body in Drosophila. *Learning & Memory* **22,** 56–63 (2015).
37. Götz, K. G. Course-Control, Metabolism and Wing Interference During Ultralong Tethered Flight in Drosophila Melanogaster. *Journal of Experimental Biology* **128,** 35–46 (1987).
38. Reiser, M. B. & Dickinson, M. H. A modular display system for insect behavioral neuroscience. *Journal of Neuroscience Methods* **167,** 127 –139 (2008).
39. Heisenberg, M., Wolf, R. & Brembs, B. Flexibility in a single behavioral variable of Drosophila. *Learn Mem* **8,** 1–10 (2001).
40. Dill, M., Wolf, R. & Heisenberg, M. Behavioral Analysis of Drosophila: Landmark Learning in the Flight Simulator. *Learning and Memory* **2,** 152–160 (1995).
41. Wolf, R. & Heisenberg, M. On the fine structure of yaw torque in visual flight orientation of Drosophila melanogaster. *J. Comp. Physiol. A* **140,** 69–80 (1980).
42. Muijres, F. T., Elzinga, M. J., Iwasaki, N. A. & Dickinson, M. H. Body saccades of Drosophila consist of stereotyped banked turns. *Journal of Experimental Biology* **218,** 864–875 (2015).
43. Turner-Evans, D. B. *et al.* The Neuroanatomical Ultrastructure and Function of a Biological Ring Attractor. *Neuron* **108,** 145–163.e10 (2020).
44. Williams, C. B. Insect Migration. *Annual Review of Entomology* **2,** 163–180 (1957).
45. Coyne, J. A. *et al.* Long-Distance Migration of Drosophila. *The American Naturalist* **119,** 589–595 (1982).
46. Wehner, R. Astronavigation in Insects. *Annual Review of Entomology* **29,** 277–298 (1984).
47. Weir, P. T. & Dickinson, M. H. Flying Drosophila orient to sky polarization. *Curr Biol* **22,** 21–27 (2012).
48. Dickinson, M. H. Death Valley, Drosophila, and the Devonian Toolkit. *Annual Review of Entomology* **59.** PMID: 24160432, 51–72 (2014).

49. Leitch, K. J., Ponce, F. V., Dickson, W. B., van Breugel, F. & Dickinson, M. H. The long-distance flight behavior of Drosophila supports an agent-based model for wind-assisted dispersal in insects. *Proceedings of the National Academy of Sciences* **118** (2021).
50. Warren, T. L., Weir, P. T. W. & Dickinson, M. H. Flying Drosophila melanogaster maintain arbitrary but stable headings relative to the angle of polarized light. *Journal of Experimental Biology* **221** (2018).
51. Haberkern, H. *et al.* Visually guided behavior and optogenetically induced learning in head-fixed flies exploring a virtual landscape. *Curr Biol* **29,** 1647–1659 (2019).
52. Mathejczyk, T. F. & Wernet, M. F. Heading choices of flying Drosophila under changing angles of polarized light. *Sci Rep* **9,** 16773 (2019).
53. Tang, S., Wolf, R., Xu, S. & Heisenberg, M. Visual Pattern Recognition in Drosophila Is Invariant for Retinal Position. *Science* **305,** 1020–1022 (2004).
54. Neuser, K., Triphan, T., Mronz, M., Poeck, B. & Strauss, R. Analysis of a spatial orientation memory in Drosophila. *Nature* **453,** 1244–1247 (2008).
55. Beetz, M. J. *et al.* State-dependent egocentric and allocentric heading representation in the monarch butterfly brain. *bioRxiv* (2021).
56. Hartmann, G. & Wehner, R. The ant's path integration system: a neural architecture. *Biological Cybernetics* **73,** 483–497 (1995).
57. Wittmann, T. & Schwegler, H. Path integration — a network model. *Biological Cybernetics* **73,** 569–575 (1995).
58. Turner-Evans, D. B. *et al.* Angular velocity integration in a fly heading circuit. *eLife* **6,** e23496 (2017).
59. Stone, T. *et al.* An Anatomically Constrained Model for Path Integration in the Bee Brain. *Current Biology* **27,** 3069–3085.e11 (2017).
60. Cope, A. J., Sabo, C., Vasilaki, E., Barron, A. B. & Marshall, J. A. R. A computational model of the integration of landmarks and motion in the insect central complex. *PLOS ONE* **12,** 1–19 (2017).
61. Kim, S. S., Rouault, H., Druckmann, S. & Jayaraman, V. Ring attractor dynamics in the Drosophila central brain. *Science* **356,** 849–853 (2017).
62. Green, J. *et al.* A neural circuit architecture for angular integration in Drosophila. *Nature* **546,** 101–106 (2017).
63. Lyu, C, Abbott, L. & Maimon, G. A neuronal circuit for vector computation builds an allocentric traveling-direction signal in the Drosophila fan-shaped body. *bioRxiv* (2020).
64. Green, J. & Maimon, G. Building a heading signal from anatomically defined neuron types in the Drosophila central complex. *Current Opinion in Neurobiology* **52.** Systems Neuroscience, 156–164 (2018).
65. Seelig, J. & Jayaraman, V. Feature detection and orientation tuning in the Drosophila central complex. *Nature* **503,** 262–266 (2013).
66. Omoto, J. J. *et al.* Visual Input to the Drosophila Central Complex by Developmentally and Functionally Distinct Neuronal Populations. *Current Biology* **27,** 1098–1110 (2017).
67. Sun, Y. *et al.* Neural signatures of dynamic stimulus selection in Drosophila. *Nat Neurosci* **20,** 1104–1113 (2017).
68. Hu, W. *et al.* Fan-Shaped Body Neurons in the Drosophila Brain Regulate Both Innate and Conditioned Nociceptive Avoidance. *Cell Reports* **24,** 1573–1584 (2018).
69. Ofstad, T. A., Zuker, C. S. & Reiser, M. B. Visual place learning in Drosophila melanogaster. *Nature* **474,** 204–207 (2011).
70. Mizunami, M., Weibrecht, J. M. & Strausfeld, N. J. Mushroom bodies of the cockroach: their participation in place memory. *J Comp Neurol* **402,** 520–537 (1998).
71. Ardin, P., Peng, F., Mangan, M., Lagogiannis, K. & Webb, B. Using an Insect Mushroom Body Circuit to Encode Route Memory in Complex Natural Environments. *PLoS Comput Biol* **12,** e1004683 (2016).
72. Collett, M. & Collett, T. S. How does the insect central complex use mushroom body output for steering? *Current Biology* **28,** R733–R734 (2018).
73. Buehlmann, C. *et al.* Mushroom Bodies Are Required for Learned Visual Navigation, but Not for Innate Visual Behavior, in Ants. *Current Biology* **30,** 3438–3443.e2 (2020).
74. Sun, X., Yue, S. & Mangan, M. A decentralised neural model explaining optimal integration of navigational strategies in insects. *Elife* **9** (2020).
75. Kamhi, J. F., Barron, A. B. & Narendra, A. Vertical Lobes of the Mushroom Bodies Are Essential for View-Based Navigation in Australian Myrmecia Ants. *Curr Biol* **30,** 3432–3437 (2020).
76. Bennett, J. E. M., Philippides, A. & Nowotny, T. Learning with reinforcement prediction errors in a model of the Drosophila mushroom body. *Nat Commun* **12,** 2569 (2021).
77. Liu, Q. *et al.* Gap junction networks in mushroom bodies participate in visual learning and memory in Drosophila. *eLife* **5,** e13238 (2016).
78. Collett, T. S. & Land, M. F. Visual control of flight behaviour in the hoverfly Syritta pipiens L. *J. Comp. Physiol.* **99,** 1–66 (1975).
79. Mongeau, J.-M. & Frye, M. A. Drosophila Spatiotemporally Integrates Visual Signals to Control Saccades. *Current Biology* **27,** 2901–2914 (2017).
80. Yttri, E. A. & Dudman, J. T. Opponent and bidirectional control of movement velocity in the basal ganglia. *Nature* **533,** 402–406 (2016).

81. Zhou, B., Hofmann, D., Pinkoviezky, I., Sober, S. J. & Nemenman, I. Chance, long tails, and inference in a non-Gaussian, Bayesian theory of vocal learning in songbirds. *Proceedings of the National Academy of Sciences* **115,** E8538–E8546 (2018).
82. Jiang, W.-C., Xu, S. & Dudman, J. Construction of a hippocampal cognitive map depends upon spatial context. *Research Gate* (Jan. 2021).
83. Reichardt, W. & Wenking, H. Optical detection and fixation of objects by fixed flying flies. *Naturwissenschaften* **56,** 424–424 (Aug. 1969).
84. Wehner, R. Spontaneous pattern preferences of Drosophila melanogaster to black areas in various parts of the visual field. *Journal of Insect Physiology* **18,** 1531–1543 (1972).
85. Horn, E. The mechanism of object fixation and its relation to spontaneous pattern preferences in Drosophila melanogaster. *Biological Cybernetics* **31,** 145–158 (1978).
86. Maimon, G., Straw, A. D. & Dickinson, M. H. A simple vision-based algorithm for decision making in flying Drosophila. *Curr Biol* **18,** 464–470 (2008).
87. Grabowska, M. J. *et al.* Innate visual preferences and behavioral flexibility in Drosophila. *Journal of Experimental Biology* **221.** jeb185918 (Dec. 2018).
88. Panser, K. *et al.* Automatic Segmentation of Drosophila Neural Compartments Using GAL4 Expression Data Reveals Novel Visual Pathways. *Curr Biol* **26,** 1943–1954 (2016).
89. Klapoetke, N. C. *et al.* Ultra-selective looming detection from radial motion opponency. *Nature* **551,** 237–241 (2017).
90. Bolton, A. D. *et al.* Elements of a stochastic 3D prediction engine in larval zebrafish prey capture. *eLife* **8** (eds Berman, G. J., Calabrese, R. L., Washbourne, P. & Combes, S. A.) e51975 (2019).
91. Demir, M., Kadakia, N., Anderson, H. D., Clark, D. A. & Emonet, T. Walking *Drosophila* navigate complex plumes using stochastic decisions biased by the timing of odor encounters. *eLife* **9** (eds Seminara, A., Calabrese, R. L., Murthy, V. N. & Celani, A.) e57524 (2020).
92. Bishop, C. M. *Pattern Recognition and Machine Learning* Second (Springer-Verlag, 2006).
93. Botvinick, M. *et al.* Reinforcement Learning, Fast and Slow. *Trends in Cognitive Sciences* **23** (2019).
94. Wang, J. X. *et al.* Learning to reinforcement learn. *arXiv:1611.05763* (2016).
95. Duan, Y. *et al.* RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning. *arXiv:1611.02779* (2016).
96. Harlow, H. The formation of learning sets. *Psychol Rev* **56** (1949).
97. Guo, A *et al.* Conditioned visual flight orientation in Drosophila: dependence on age, practice, and diet. *Learning & Memory* **3,** 49–59 (1996).
98. JW, W., AM, W., J, F., LB, V. & R, A. Two-photon calcium imaging reveals an odor-evoked map of activity in the fly brain. *Cell* **112,** 271–282 (2003).
99. Seelig, J. D. *et al.* Two-photon calcium imaging from head-fixed Drosophila during optomotor walking behavior. *Nature Methods* **7,** 535–540 (2010).
100. Hubert, M. & Vandervieren, E. An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis* **52,** 5186–5201 (2008).
101. Shiozaki, H. & Kazama, H. Parallel encoding of recent visual experience and self-motion during navigation in Drosophila. *Nat Neurosci* **20,** 1395–1403 (2017).
102. B.j., H. *et al.* A visual pathway for skylight polarization processing in Drosophila. *eLife* **10,** e63225 (2021).
103. Heinze, S. & Reppert, S. M. Sun Compass Integration of Skylight Cues in Migratory Monarch Butterflies. *Neuron* **69,** 345–358 (2011).
104. Vitzthum, H., Müller, M. & Homberg, U. Neurons of the Central Complex of the Locust Schistocerca gregaria are Sensitive to Polarized Light. *Journal of Neuroscience* **22,** 1114–1125 (2002).
105. Okubo, T., Patella, P, D'Alessandro, I & Wilson, R. A Neural Network for Wind-Guided Compass Navigation. *Neuron* **107,** 924–940 (5 2020).
106. Hanesch, U., Fischbach, K.-F. & Heisenberg, M. Neuronal architecture of the central complex in Drosophila melanogaster. *Cell and Tissue Research* **257,** 343–366 (1989).
107. Skaggs, W., Knierim, J., Kudrimoti, H. & McNaughton, B. A model of the neural basis of the rat's sense of direction. *Adv Neural Inf Process Syst* **7,** 173–180.
108. Kuntz, S., Poeck, B. & Strauss, R. Visual Working Memory Requires Permissive and Instructive NO/cGMP Signaling at Presynapses in the Drosophila Central Brain. *Current Biology* **27,** 613–623 (2017).
109. Liang, X. *et al.* Morning and Evening Circadian Pacemakers Independently Drive Premotor Centers via a Specific Dopamine Relay. *Neuron* **102,** 843–857.e4 (2019).
110. Franconville, R., Beron, C. & Jayaraman, V. Building a functional connectome of the *Drosophila* central complex. *eLife* **7** (eds VijayRaghavan, K, Scott, K. & Heinze, S.) e37017 (2018).
111. Scheffer, L. K. *et al.* A connectome and analysis of the adult *Drosophila* central brain. *eLife* **9** (eds Marder, E., Eisen, M. B., Pipkin, J. & Doe, C. Q.) e57443 (2020).
112. Lin, C. Y. *et al.* A comprehensive wiring diagram of the protocerebral bridge for visual information processing in the Drosophila brain. *Cell Rep* **3,** 1739–1753 (2013).
113. Wolff, T., Iyer, N. A. & Rubin, G. M. Neuroarchitecture and neuroanatomy of the Drosophila central complex: A GAL4-based dissection of protocerebral bridge neurons and circuits. *J Comp Neurol* **523,** 997–1037 (2015).

114. Currier, T. A., Matheson, A. M. & Nagel, K. I. Encoding and control of orientation to airflow by a set of Drosophila fan-shaped body neurons. *Elife* **9** (2020).
115. Lu, J *et al.* Transforming representations of movement from body- to world-centric space. *bioRxiv* (2020).
116. Matheson, A. M. *et al.* A neural circuit for wind-guided olfactory navigation. *bioRxiv* (2021).
117. Müller, J., Nawrot, M., Menzel, R. & Landgraf, T. A neural network model for familiarity and context learning during honeybee foraging flights. *Biological Cybernetics* **112,** 113–126 (2018).
118. Zhu, L., Mangan, M. & Webb, B. Spatio-temporal Memory for Navigation in a Mushroom Body Model. *bioRxiv* (2020).
119. Claridge-Chang, A. *et al.* Writing Memories with Light-Addressable Reinforcement Circuitry. *Cell* **139,** 405–415 (2009).
120. Aso, Y. & Rubin, G. M. Dopaminergic neurons write and update memories with cell-type-specific rules. *eLife* **5** (ed Luo, L.) e16135 (2016).
121. Cohn, R., Morantte, I. & Ruta, V. Coordinated and Compartmentalized Neuromodulation Shapes Sensory Processing in Drosophila. *Cell* **163,** 1742–1755 (2015).
122. Cognigni, P., Felsenberg, J. & Waddell, S. Do the right thing: neural network mechanisms of memory formation, expression and update in Drosophila. *Current Opinion in Neurobiology* **49.** Neurobiology of Behavior, 51–58 (2018).
123. Siju, K. *et al.* Valence and State-Dependent Population Coding in Dopaminergic Neurons in the Fly Mushroom Body. *Current Biology* **30,** 2104–2115.e4 (2020).
124. Weir, P. T., Schnell, B. & Dickinson, M. H. Central complex neurons exhibit behaviorally gated responses to visual motion in Drosophila. *Journal of Neurophysiology* **111.** PMID: 24108792, 62–71 (2014).
125. Hsu, C. T. & Bhandawat, V. Organization of descending neurons in Drosophila melanogaster. *Scientific Reports* **6,** 20259 (2016).
126. Namiki, S., Dickinson, M. H., Wong, A. M., Korff, W. & Card, G. M. The functional organization of descending sensory-motor pathways in *Drosophila*. *eLife* **7** (ed Scott, K.) e34272 (2018).
127. Cande, J. *et al.* Optogenetic dissection of descending behavioral control in *Drosophila*. *eLife* **7** (ed Scott, K.) e34275 (2018).

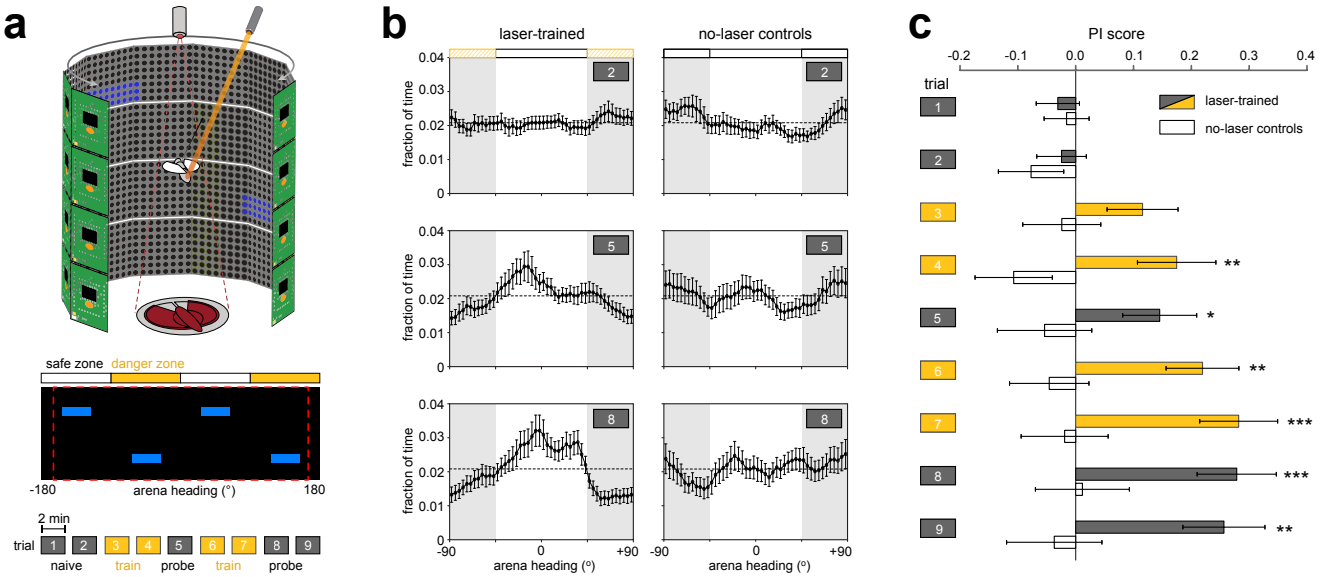

Figure 1: Flies operantly learn to avoid heat punishment. a) Upper panel: Schematic of flight simulator and LED arena. Middle panel: The arena is divided into four quadrants; one set of opposing quadrants contains horizontal bars at a high elevation, and the other set of quadrants contains horizontal bars at a low elevation. During training trials, one set of quadrants ('danger zone') is paired with an aversive heat punishment (orange bars), while the other set of quadrants ('safe zone') remain unpunished (white bars). The red dashed box indicates the span of the visual arena. Lower panel: training protocol. During training trials, laser punishment is delivered whenever the fly's heading falls within the danger zone. During naive and probe trials, no punishment is delivered. Each trial lasts 2 min. **b)** Average residency (expressed as a fraction of total time) spent at different headings in the arena, measured in laser-trained flies (left column; $n = 44$) and no-laser control flies (right column; $n = 40$) and aligned to the center of the safe zone. Note that arena headings are folded from $\pm 180°$ to $\pm 90°$ based on the symmetry of the visual scene. Top, middle, and bottom rows: trials 2, 5, and 8, respectively. Error bars: mean +/- standard error. **c)** Average Performance Index (PI) scores for laser-trained and no-laser control flies. Error bars: mean +/- standard error. Significance: Wilcoxon rank sum test ($^{*}p \leq 0.05$; $^{**}p \leq 0.01$; $^{***}p \leq 0.001$).

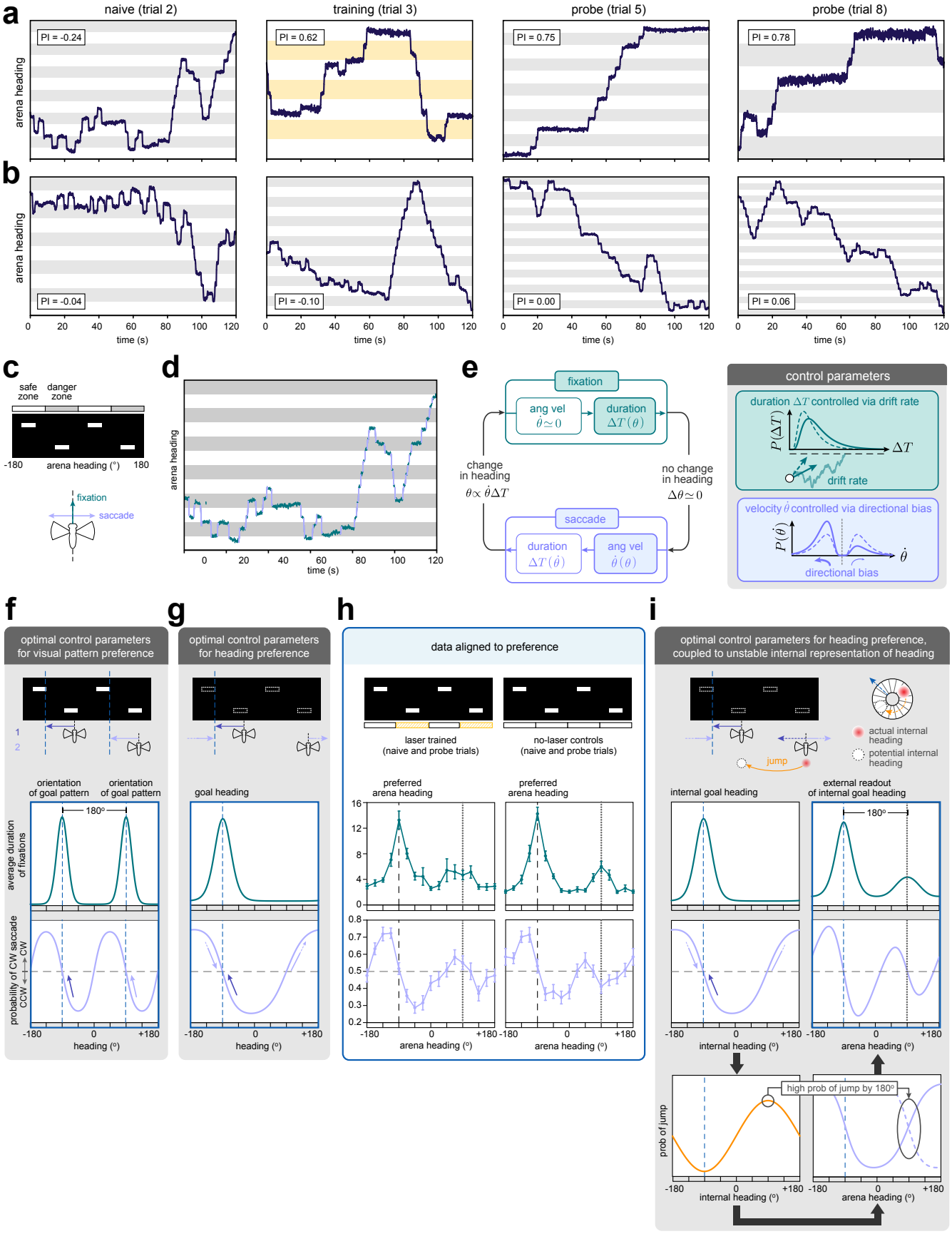

Figure 2: An inferred behavioral policy captures conditioned and unconditioned behavior. a) Example heading trajectories and PI scores from a single fly that underwent laser training. Trajectories were unwrapped to reveal the overall structure in the behavior (Methods). Gray and yellow bars indicate headings within the danger zone that are punished (yellow bars; training trial), or will be/have been punished (gray bars; naive/probe trials). **b)** Same as (**a**), for a no-laser control fly that did not undergo laser training. **c)** Flies exhibit periods of straight flight in which they maintain a constant arena heading (fixations, green arrow), punctuated by abrupt turns that lead to changes in heading (saccades; purple arrows). **d)** Same behavioral trace as shown in the left panel of (**a**), segmented into fixations (green) and saccades (purple). **e)** A behavioral policy, which can be used to generate behavior in model flies, consists of alternating modes of fixations and saccades. Each mode is specified by an angular velocity $\dot{\theta}$ that is maintained over a duration of time ΔT. Model flies can control both the duration of fixations (maintained at near-zero angular velocity; green shaded boxes) and the angular velocity of saccades (maintained for a velocity-dependent duration; purple shaded boxes) as a function of heading θ. **f-g)** Optimal control parameters for two different models that are trained via a reinforcement learning algorithm using the behavioral policy in **(e)** to maintain a preference for a specific visual pattern (**f**) or for a goal heading (**g**) (Methods). Note that the optimal control parameters specify the drift rate of fixations; the resulting average duration of fixations scales inversely with this drift rate. **f)** Control parameters optimized for maintaining a preferred visual pattern, or 'goal pattern' (indicated by the blue dashed vertical lines), result in a behavioral policy that is structured as a function of the difference between the orientation of the current visual pattern and that of the goal pattern. This policy manifests in a symmetric bimodality in both the duration of fixations (green curve) and the direction of saccades (purple curve), centered about the two headings corresponding to the symmetric orientations of the preferred pattern. **g)** Control parameters optimized for maintaining a goal heading (indicated by the blue dashed vertical line) result in a behavioral policy that is structured as a function of the difference between the current and goal headings. This policy manifests in a unimodality in both the duration of fixations (green curve) and the direction of saccades (purple curve), centered about the goal heading. Note the difference in actions predicted by the two policies in (**f**) and (**g**), as schematized at the top of each panel: in scenario 1, when the model fly maintains a heading near -45°, both policies predict that a CCW turn is more likely (dark purple arrows). In contrast, in scenario 2, when the model fly maintains a heading near 135°, the visual-pattern-based policy predicts that a CCW turn is more likely, while the goal-heading-based policy predicts that a CW turn is more likely (light purple arrows). **h)** Average duration of fixations (upper row) and direction of saccades (lower row) measured in laser-trained flies (left column; $n = 44$) and no-laser control flies (right column; $n = 40$), aligned to preferred arena heading (vertical dashed line; vertical dotted line marks the 180°-symmetric heading). Error bars: mean +/- standard error. **i)** When the control parameters shown in (**g**) are coupled to an unstable internal representation of heading (orange curve in lower left), the resulting behavioral readout (right column) shows an asymmetric bimodality in both the duration of fixations and the direction of saccades, with the larger of the two peaks centered about the goal heading and the weaker of the two peaks centered about the 180°-symmetric heading. This bimodality arises because the internal representation can jump between orientations corresponding to symmetric views of the visual scene (upper schematic), and thus leads to a partial "copying-over" of the behavioral policy at symmetric arena headings (see Methods and SI Fig S4 for more details).

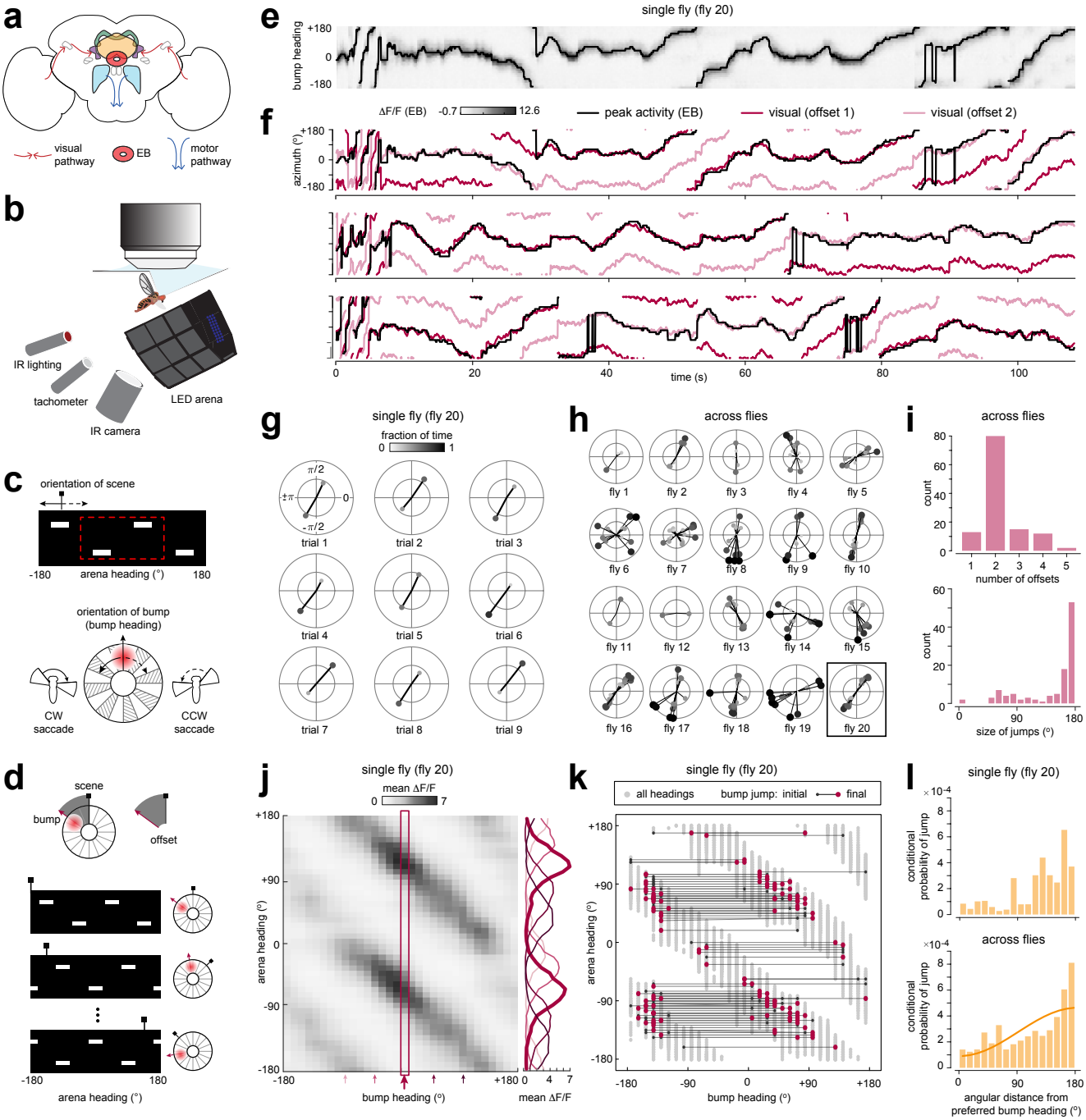

Figure 3: Symmetric scenes induce instabilities in the offset of the heading bump. **a)** Upper: Schematic of central complex (CX), highlighting pathways to the ellipsoid body (EB) and motor pathways leaving the CX. **b)** Schematic of two-photon calcium imaging setup for tethered flies flying inside an LED arena. **c)** An internal representation of heading is maintained as a bump of activity (red) in the EB. The bump moves in the opposite direction that the fly turns to track the fly's orientation relative to the visual scene. The red dashed box indicates the span of the LED arena positioned under the two-photon microscope. **d)** We measure the offset as the angular difference between the heading bump and the visual scene. **e)** EPG neuron activity (heatmap) and location of peak activity (black line) during closed-loop tethered flight with a symmetric visual scene. **f)** The peak activity (black) jumps between two different offsets (pink and maroon) that correspond to symmetric views of the scene. Shown for trials 1-3. **g)** Offsets (measured as the circular distance between the peak activity and a reference orientation of the visual scene) for all 9 trials of the same fly shown in (**e,f**). Size, color, and radial distance of marker from the origin indicate fraction of time spent at a given offset. **h)** Same as (**d**), but collapsed across trials for 20 flies. **i)** Upper: Number of unique offsets, aggregated across flies and trials. Lower: Angular separation between the two dominant offsets, aggregated across flies and trials. **j)** Main panel: Average $\Delta F/F$ of different wedges in the EB ("bump heading") as a function of arena heading, averaged across all 9 trials for the fly shown in panels **e-g**. Note that this representation differs from that of panel (**f**), where we displayed the azimuthal scene orientation. Here, to maintain consistency with the other analyses in this study, we display the arena heading, which differs from the scene orientation by a sign flip. Right panel: Tuning curve for different EB wedges (marked by arrows in main panel). **k)** Bump jumps as a function of location in the EB, collapsed across all 9 trials for the fly in (**e,f**), shown for jumps that exceeded a threshold of 135°. **l)** Probability of a bump jump as a function of angular distance from the preferred bump heading, shown for a single fly (upper panel) and aggregated across flies (lower panel). For each fly, the conditional probability was estimated as the fraction of all visits to a given EB wedge (accumulated across trials) for which the angular size of the bump jump exceeded 135°. Bars: average probability across flies. Solid line: best-fitting cosine curve.

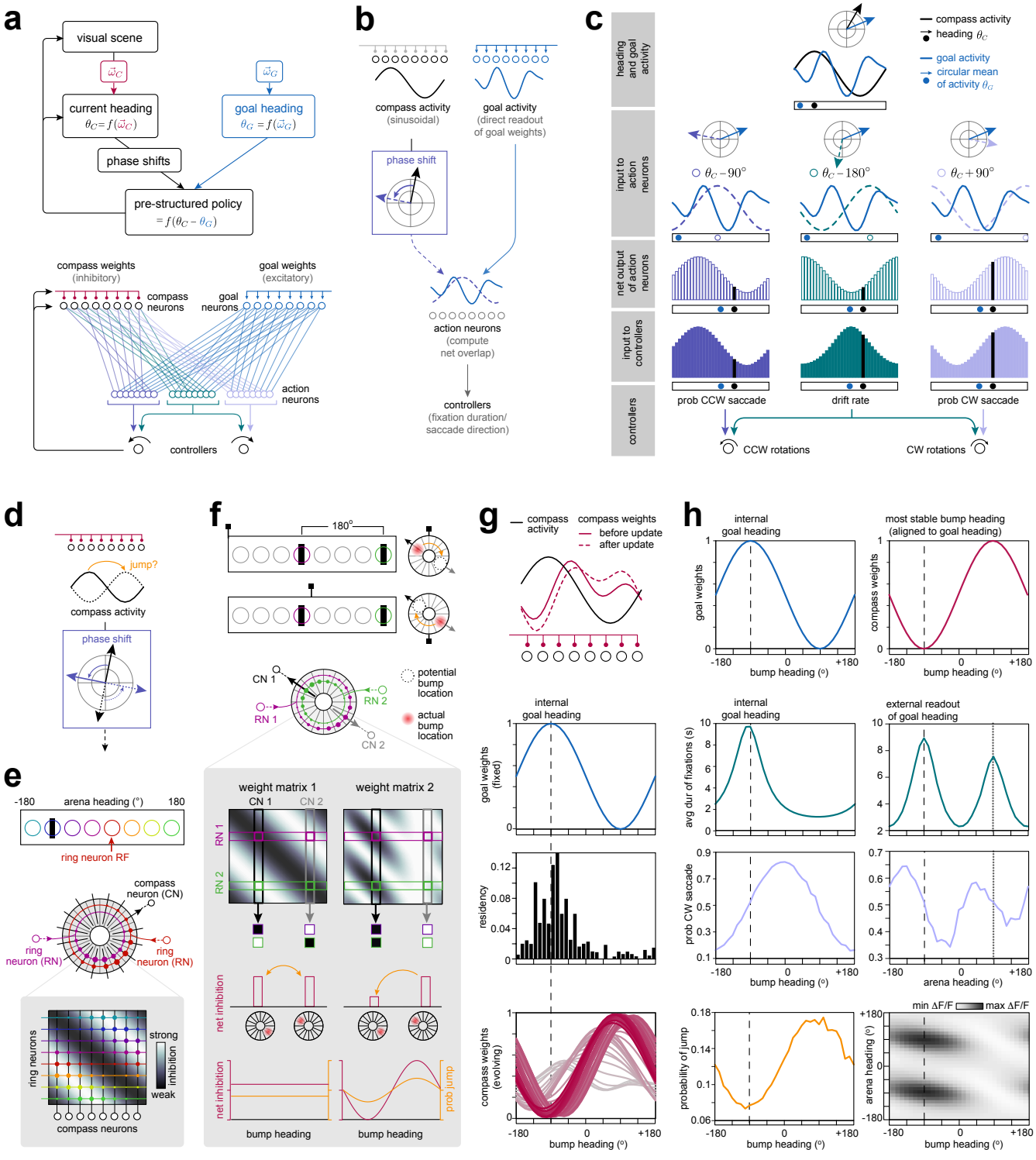

Figure 4: A simple circuit model reproduces observed behavior. a) Basic logic of circuit model (upper panel) and network implementation (lower panel). The current heading θ_C is maintained in a population of compass neurons. The stability of the current heading is determined by the properties of the visual scene and captured by a set of inhibitory compass weights $\vec{\omega}_C$. The goal heading θ_G is maintained in the activity of a population of goal neurons, and is determined by a set of excitatory goal weights $\vec{\omega}_G$. A prestructured internal policy is maintained by a population of action neurons and downstream controllers, which use the overlap between the goal heading and phase-shifted versions of the current heading to specify actions as a function of the difference between current and goal headings. **b)** Action neurons compute the net overlap between the goal activity and a phase-shifted version of the compass activity, and send their output to controllers that govern the duration of fixations and directionality of saccades; this operation ensures that the behavior is structured with respect to the angular difference between current and goal headings. **c)** First row: Compass and goal neurons carry activity about the fly's current and goal headings (black and blue curves, respectively). The circular means of the heading and goal activity profiles are marked by black and blue dots below the panel, and by polar plots above the panel. Here, we consider a multi-peaked goal activity profile to illustrate how this circuit architecture can maintain a prestructured policy even in the absence of sinusoidal inputs. As we will later show, learning will drive the goal activity profile to be sinuoidal. Second row: Three populations of action neurons receive phase-shifted compass activity (dashed lines) and goal activity (solid blue lines); each neuron in these populations multiplicatively combines these inputs. Third row: The summed output of each action neuron population varies as a function of the fly's current heading (black dot) relative to the circular mean of the goal activity (blue dot; aligned to be at the center of each plot); this circular mean thus defines the goal heading. Filled bars correspond to the value at the heading shown in the first row. Fourth and fifth row: Action neurons project to descending neurons that initiate CW/CCW turns; the net motor drive controls the probability of initiating a CW or CCW saccade (purple) and maintaining a given duration of fixation (green) depending on the fly's current heading (black dot) relative to the goal heading (blue dot). **d)** Compass weights capture the net inhibition onto compass neurons, and determine whether the heading bump will jump. If the bump jumps, this new bump heading will change all downstream computations that rely on phase-shifted versions of the current heading. **e)** Visual ring neurons have receptive fields that tile visual space (upper). These ring neurons convey sensory input to the EB (middle), where they make all-to-all inhibitory synapses onto compass neurons (lower). Inhibitory Hebbian-like plasticity between coactive ring neurons and compass neurons leads to a structured pattern of inhibition (heatmap). We summarize this pattern of inhibition as a single vector of weights onto compass neurons (as schematized in (**d**)). **f)** Upper panels: In a symmetric scene, the same ring neurons will be active for two different views of the scene (and thus for two different EB locations of the heading bump). In this case, the profile of inhibition determines which of these two EB locations is more stable, and thus which location the bump is more likely to occupy. Lower left column: If the mapping is consistent across the EB (that is, if each compass neuron receives a shifted but otherwise identical weight profile from its input ring neurons, as shown in **e**), the net inhibition from active ring neurons onto compass neurons is the same for both orientations of the bump (red bars), and thus the bump will be equally likely to jump in either direction (orange arrow). When generalized across all headings, the relative difference in net inhibition (red curve, lower left) between any two headings separated by 180° determines the probability that the HD bump will jump (orange curve, lower left). Lower right column: If the mapping is not consistent across the EB, the net inhibition from active ring neurons onto compass neurons will differ for the two orientations of the bump (red bars), and the bump will be more likely to jump in one direction than the other (orange arrow). This will lead to a nonuniform probability of jumps across all headings (orange curve, lower right). **g)** Inhibitory Hebbian-like plasticity updates the compass weights (top panel, solid and dashed red lines) in proportion to local compass neuron activity (black line); note that we used an artificially high learning rate to illustrate this update (Methods). Given a fixed profile of goal weights (second panel), the behavioral policy derived from these goal weights will drive the heading bump to sample some regions of the EB more than others (third panel), which in turn drives the compass weights to develop a sinusoidal structure whose most stable heading is aligned with the goal heading (bottom panel). **h)** We use a fixed set of goal weights (blue, upper left) and compass weights (red, upper right) to simulate the behavior of model flies in the absence of training, assuming that the most stable bump heading is aligned to the internal goal heading (as shown in (**g**) and indicated by the alignment of the dashed vertical lines). When conditioned on the current bump heading (left column, second and third rows), the average duration of fixations and directionality of saccades exhibits a unimodal structure. Because the bump heading is unstable, and because this instability is structured as a function of heading (lower left), this internal policy manifests in a bimodal external readout (right column, second and third rows), and a bimodal tuning of compass neurons to different arena headings (bottom right). Note that here, in contrast to the model shown in Fig 2i, we explicitly model saccades as moving the fly and the bump in opposite directions, and thus internal and external directionality of saccades (third row) are inverted with respect to one another about $P_{CW} = 0.5$.

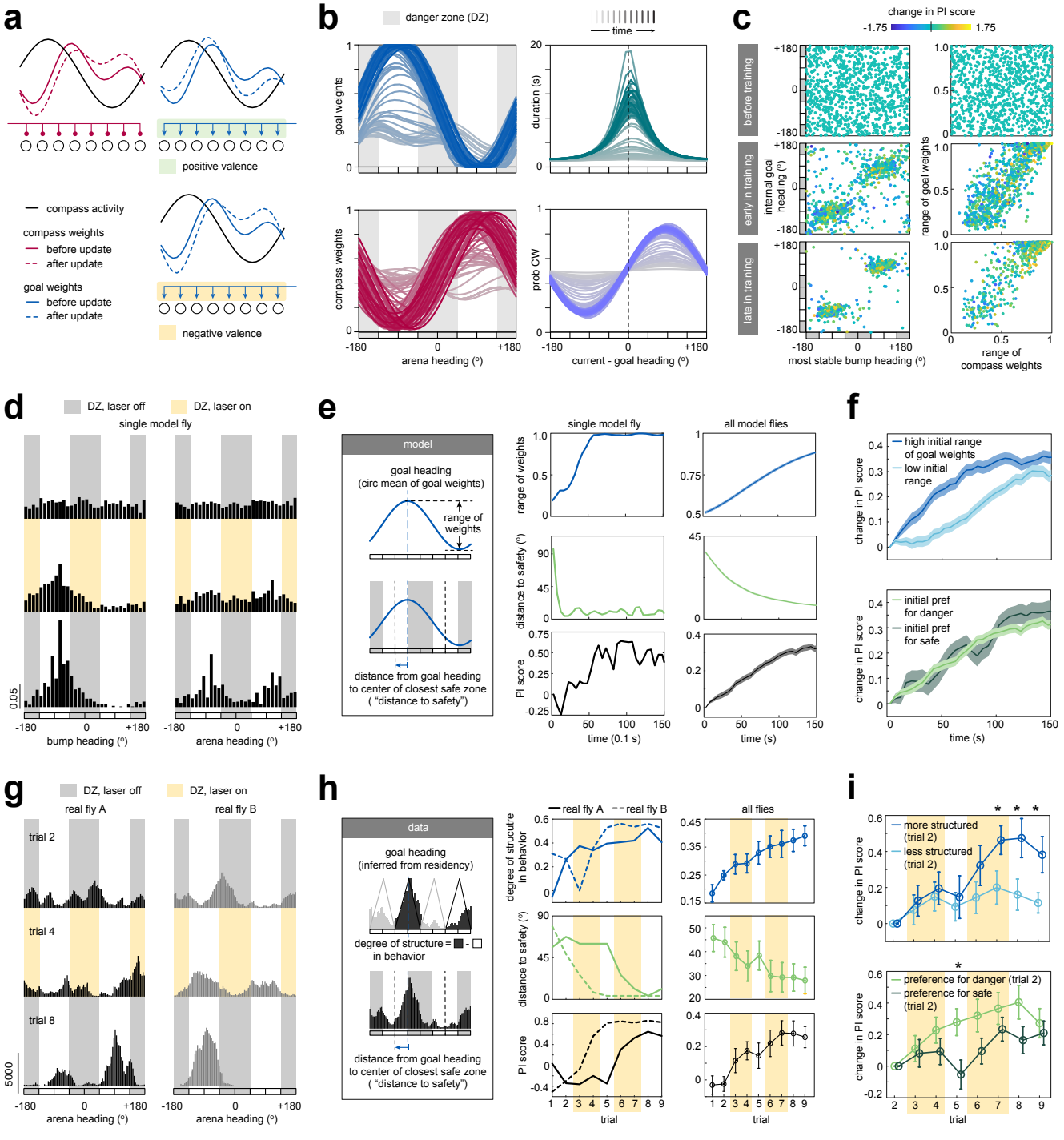

Figure 5: Heading and goal weights influence performance. a) Hebbian-like plasticity updates the goal weights (solid and dashed blue lines, before and after the update, respectively) in proportion to the compass neuron activity (black curve), depending on the valence (positive, upper row; negative, lower row). As in Fig 4g, we used an artifically high learning rate for illustration. **b)** Example learning trajectory from a single model fly; color saturation indicates time during training (darker colors indicate later times); gray bars indicate danger zones (DZs). Weights onto compass (red) and goal (blue) neurons are initialized with weak values centered on different headings, which together only weakly drive fixations (green) and saccades (purple). Over the course of training, the compass and goal weights co-stabilize in one of the two safe zones (here, the lefthand safe zone), such that the point of weakest inhibition onto the compass neurons (corresponding to the most stable heading) is aligned with the circular mean of the goal weights (corresponding to the goal heading). As the weights co-stabilize, their range increases, which leads to an increased motor drive (SI Fig S7). **c)** Variability in the location (left column) and range (right column) of compass and goal weights across 1000 different model initializations, shown before training (upper row), one-quarter of the way through training (middle row), and three-quarters of the way through training (lower row). Color code indicates the change in PI score at these successive timepoints in training. The location and range of compass and goal weights are randomly initialized prior to training (upper row; Methods), and they co-stabilize over the course of training. This is marked by the emergence of diagonal structure (middle row) and the shifting of the goal heading toward one of the two safe zones (lower left panel). **d)** Residency histograms of the bump heading (left column) and arena heading (right column) from the simulation shown in (**b**). Initially, the bump and the fly sample heading space uniformly. As the goal heading stabilizes in the left-hand safe zone, the fly's behavior will increasingly drive the bump to this location; as the most-stable heading stabilizes in this same location, the bump will be increasingly less likely to jump from this location. This behavior emerges even as the fly samples both safe zones (right column), because the bump will increasingly tend to jump back to the left-hand safe zone when the fly is sampling the right-hand safe zone. **e)** Over the course of training, the range of the goal weights (and thus the degree of structure in the behavior) increases (upper) and the goal heading shifts towards the center of the safe zone (middle), which together lead to an increase in PI scores (lower). This is consistent within individual flies (left column; shown for the same model fly as in **b**) and across model flies (right column; shown for the same simulations as in **c**). **f)** When the model flies (taken from the simulations shown in (**c**)) are separated based on the initial range of their goal weights (and thus the initial degree of structure in their behavior; upper panel) or their initial heading preference (lower panel), flies that initially exhibited more structured behavior showed larger changes in PI scores over the course of training, whereas initial preference did not impact average performance. Shaded regions: mean +/- standard error. **g)** Residency histograms from two real flies that showed large and positive changes in PI scores after training. Fly A initially exhibited a weak and distributed behavioral preference for different headings; over time, it developed a strong behavioral preference for both safe zones. Fly B initially exhibited a strong preference for the region straddling the safe and danger zones, and shifted this to the left-hand safe zone. **h)** Over the course of training, the two individual flies in (**g**) showed an increase in the degree of structure in their behavior (upper left), a decrease in the distance between their inferred goal heading and the center of the closest safe zone (middle left), and an increase in PI scores (lower left). The same trends were observed across the population (right column). Error bars: mean +/- standard error. **i)** Upper: Flies with more structured behavior showed consistently larger changes in PI scores after training relative to flies with less structured behavior (upper). In comparison, flies did not significantly differ in their final performance when separated by initial preference (lower). Error bars: mean +/- standard error. Significance: Wilcoxon rank sum test ($^{*}p \leq 0.05$).

SUPPLEMENTAL FIGURES

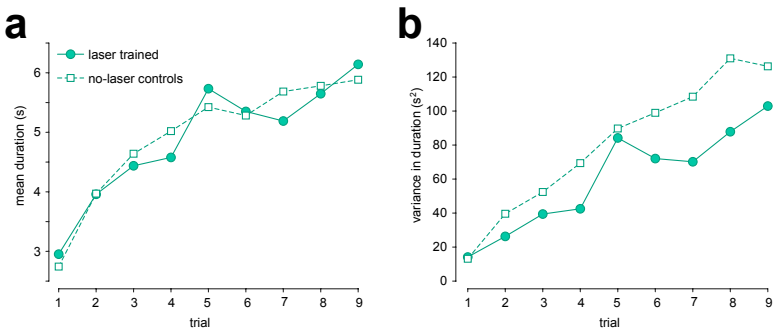

Figure S1: Average fixation duration increases over time. Average (**a**) and variance (**b**) in the distribution of fixation durations, measured across flies for a given trial and estimated via an Inverse Gaussian fit (see *Analysis Methods: Characterizing fixation properties* for fitting procedure).

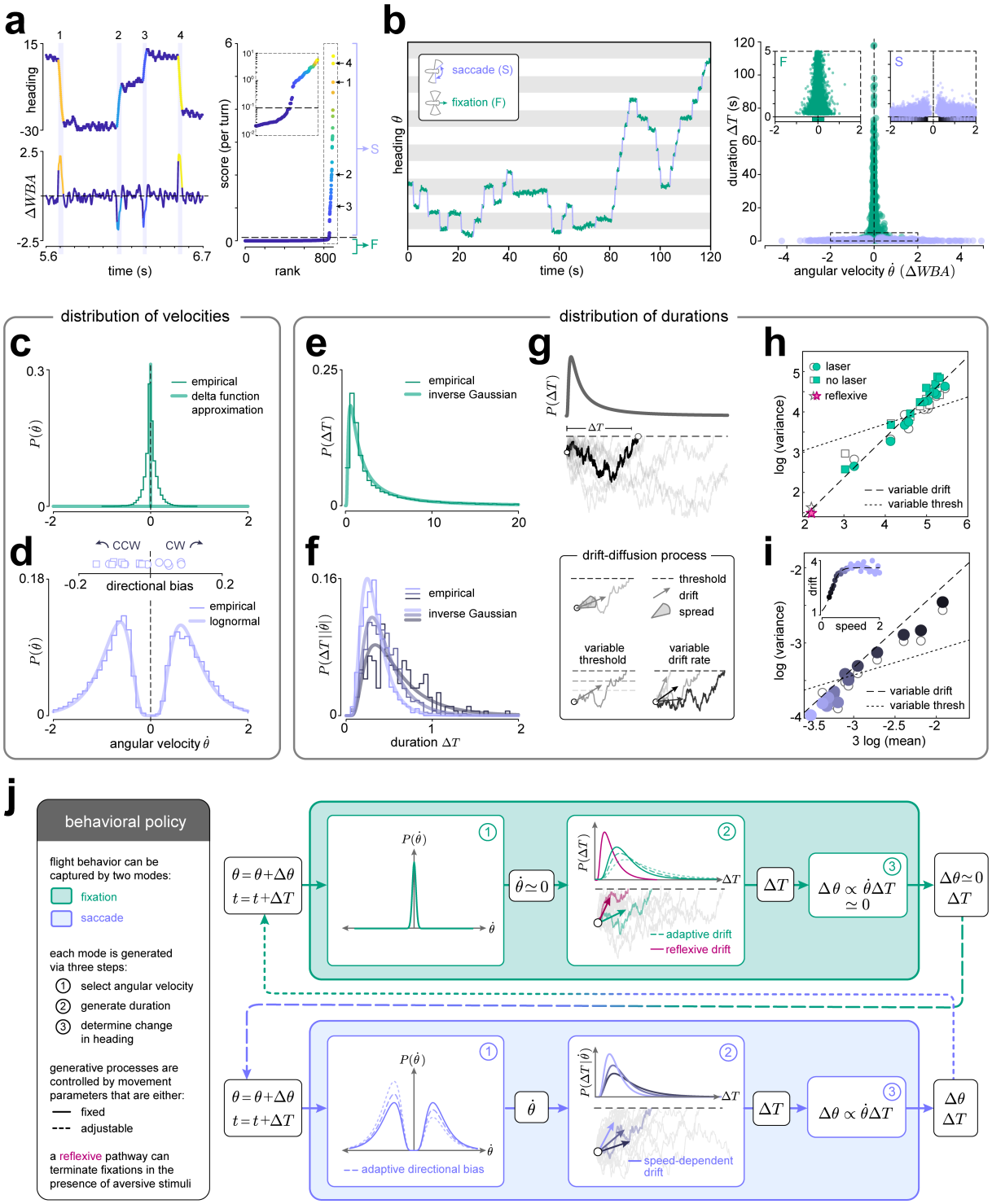

Figure S2: Inferring a behavioral policy. a) Example detection of saccades from a portion of the heading trajectory shown in (**b**) (and also shown in Fig 2a). Left column: example saccades, numbered and colored according to score (right column; explained below). Right column: Individual turns are scored based on changes in heading and wing beat amplitude (see *Analysis Methods: Partitioning behavior into fixations and saccades*). Turns with scores above a threshold (dashed line) are defined as saccades ('S'). Periods between saccades are defined as fixations ('F'). Inset: same data, shown using a log scale on the score. **b)** Left: Heading trajectory partitioned into fixations (green) and saccades (purple). Right: Distribution of duration Δt and angular velocity $\dot{\theta}$ of individual fixations and saccades, accumulated over 9 trials across laser-trained ($n = 44$) and no-laser control ($n = 40$) flies. For each event (fixation or saccade), Δt measures the entire duration of the event; $\dot{\theta}$ is measured as the change in wing beat amplitude, averaged across the event. Insets: distribution events with short durations and low angular velocity; colored bars along horizontal axis indicate 95% confidence intervals. **c)** Empirical distribution of fixation velocities (using data in (**b**)), approximated as a delta function. **d)** Empirical distributions of clockwise and counter clockwise saccade velocities (using data in (**b**)), and best-fitting lognormal approximations. Inset: directional bias of saccades; each marker shows the bias measured across flies within individual trials. **e)** Empirical distribution of fixation durations (using data in (**b**)), and best-fitting inverse Gaussian approximation. **f)** Empirical distribution of saccade durations conditioned on different saccade velocities (using data in (**b**)), and best-fitting inverse Gaussian approximations. **g)** Illustration of drift-diffusion (DD) process for generating inverse Gaussian (IG) distribution of durations. For a given choice of the mean drift rate, spread, and threshold of the diffusion process, the time of first threshold crossing ('first passage' time) is IG-distributed. **h)** Properties of the within-trial across-fly distributions of fixation durations, measured from the best-fitting IG distributions (green filled markers) and estimated empirically (gray open markers), compared against a model in which the variability in these same properties arises from changes in either the drift rate (dashed line) or threshold (dotted line) of a DD process (see schematic in **g**). Star: distribution of fixation durations estimated from fixations made within the danger zone during the first 60 s of the first training trial. **i)** Same as **h**, but shown for properties of the across-trial across-fly distributions of saccade durations conditioned on different angular speeds. Inset: the drift rate of the best-fitting DD process is nonlinearly related to the average angular speed of saccades, and can be fit with a sigmoidal function (dashed line; see *Analysis Methods: Characterizing saccade properties*). **j)** Left: ingredients of behavioral policy. Right: schematic of policy consisting of transitions between fixations (green box) and saccades (purple box). Fixations are initiated with zero angular velocity (F1), and the duration of a given fixation is generated on-line via a DD process with an adaptive drift rate (F2). When receiving punishment, this process can be short-circuited by a DD process with a reflexive drift rate. The fixation is terminated when either the adaptive or reflexive DD process first crosses a fixed threshold, leading to a change in time but no change in heading (F3). Following the termination of a fixation, a saccade is initiated. The angular velocity of the saccade is sampled from a lognormal distribution with adaptive directional bias (S1), and the duration of the saccade is generated via a DD process whose drift rate depends on this angular velocity (S2). The saccade is terminated when the DD process first crosses a fixed threshold, leading to a change in both time and heading (S3). Following the termination of a saccade, a fixation is initiated, and the process repeats.

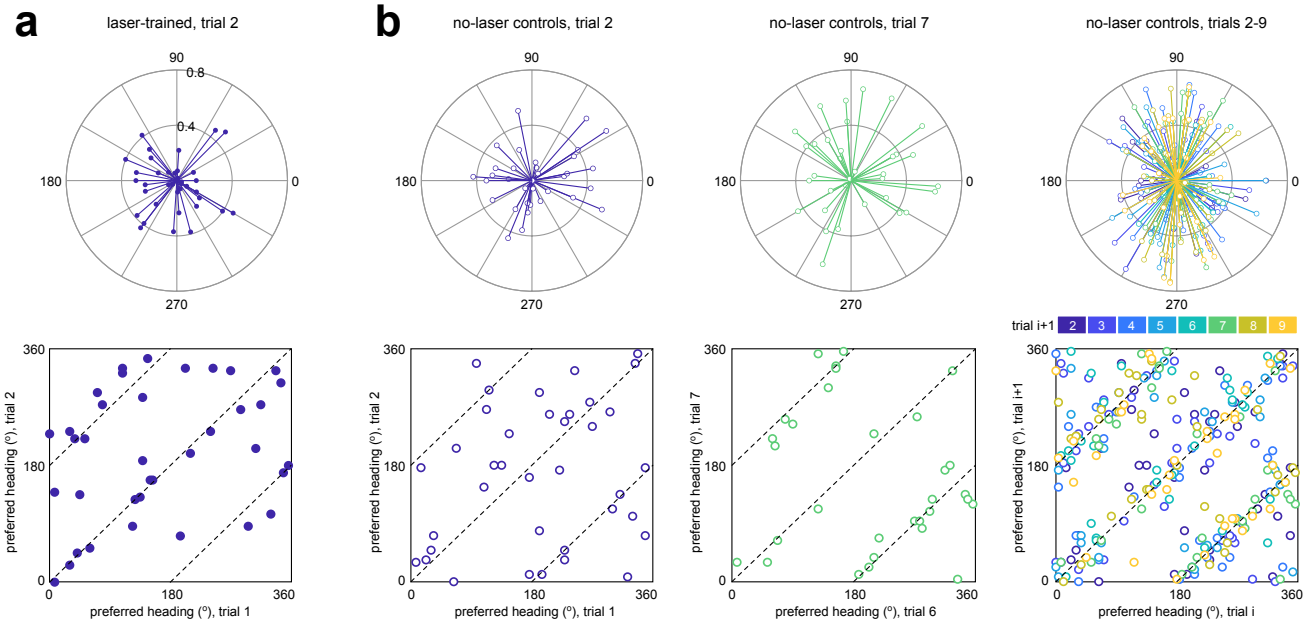

Figure S3: Flies display variability in heading preferences. a) Upper: angular location and strength of heading preferences for laser-trained flies, measured in trial 2 (see *Analysis Methods: Aligning to individual preference* and *Analysis Methods: Measuring the degree of structure in the behavior* for measurements of location and strength, respectively). Lower: comparison of angular location of heading preferences between trials 1 and 2. Due to the symmetry of the visual scene, we would expect preferences between successive trials to be similar up to a shift of $\pm 180°$ (dashed lines). **b)** Upper row: angular location and strength of heading preferences for no-laser control flies. Lower: comparison of angular location of heading preferences between successive trials.

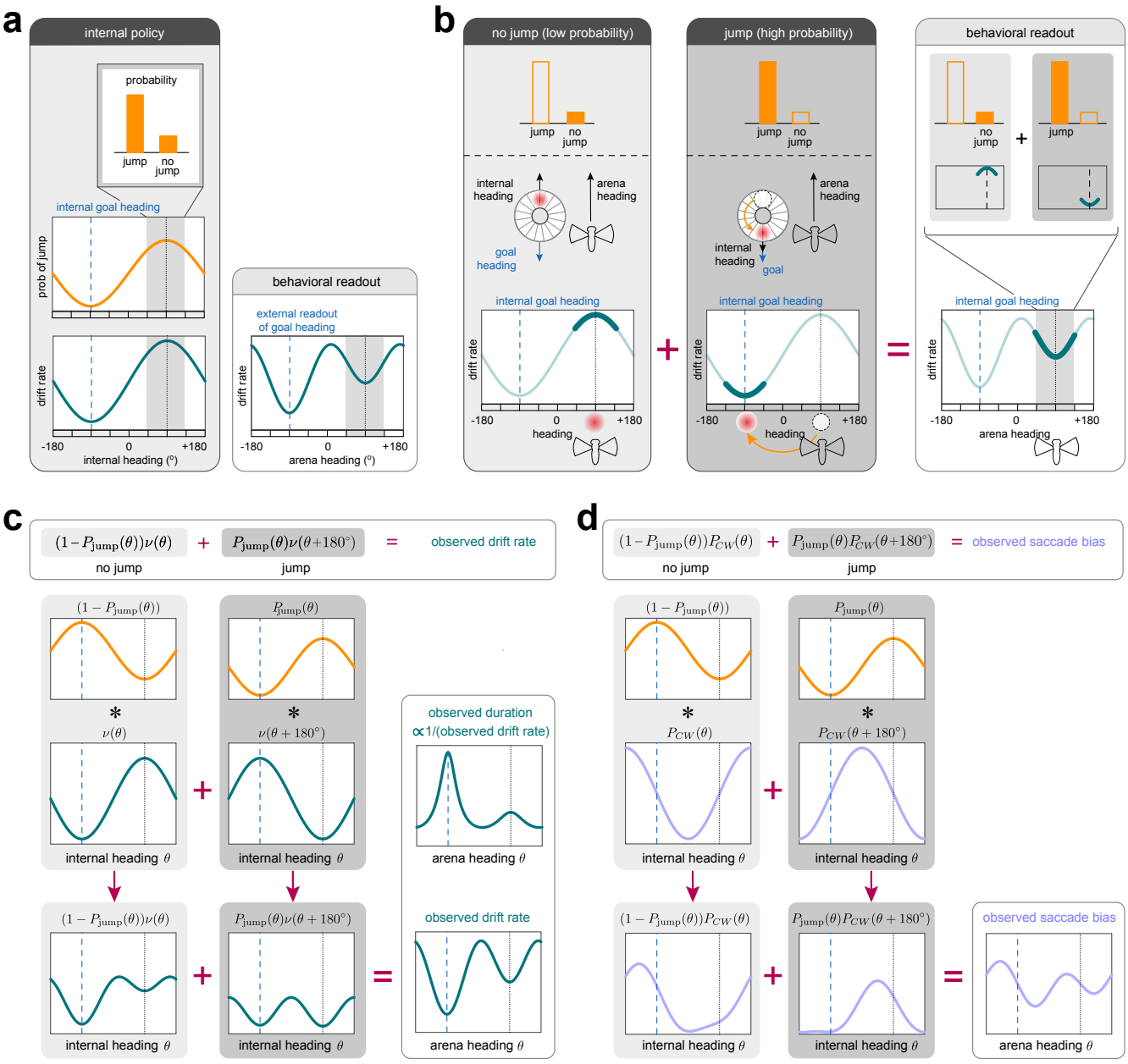

Figure S4: Structured heading instability leads to bimodality in behavioral readout. a) Our model fly's internal policy specifies the drift rate of fixations (green curve, left box) and the directional bias of saccades (not shown) as a function of the fly's internal heading relative to a goal heading (blue dashed line). We consider a scenario in which this internal heading is unstable, and can probabilistically jump between different symmetric views of the visual scene (orange curve, left box). In the two-fold symmetric scene used here, this corresponds to a jump of 180°. The observed behavioral readout (right box) is a weighted sum of the two scenarios in which the bump does or does not jump. **b)** Consider a scenario in which the fly is at an arena heading of $+90^\circ$, and the bump heading (red circle) is at the same orientation (corresponding to the gray highlighted region in (**a**)). At this particular bump heading, there is a high probability that the bump will jump, and a low probability that the bump will not jump. If the bump does not jump, the fly's behavior will be determined by the drift rate at $+90^\circ$ (green highlighted portion of the drift rate curve in left box). If the bump does jump (orange arrow), it will jump by 180° (based on the two-fold symmetry of the scene), and the fly's behavior will be determined by the drift rate at -90° (green highlighted portion of the drift rate curve in middle box). In both cases, the fly is at the arena heading of $+90^\circ$. Its average behavior at this arena heading will be the weighted sum of a high drift rate with low probability (corresponding to the scenario when the bump did not jump), and a low drift rate with high probability (corresponding to the scenario when the bump did jump). **c)** When generalized across all possible arena headings, the observed drift rate will be a weighted sum of the two scenarios in which the bump did and did not jump. To illustrate the resulting drift rate across all headings θ, we define the internally-generated drift rate as $\nu(\theta)$, and the internally-generated probability of a jump as $P_{\text{jump}}(\theta)$. We first illustrate the scenario in which the bump does not jump (left column), which happens with probability $(1 - P_{\text{jump}}(\theta))$ (upper panel, left column). In this case, the fly's behavior at an arena heading θ is determined by the drift rate at that same heading, given by $\nu(\theta)$ (middle panel, left column). The point-wise product between these curves (lower panel, left column) gives the first contribution to the observed drift rate. We next illustrate the scenario in which the bump does jump (middle column), which happens with probability $P_{\text{jump}}(\theta)$ (upper panel, middle column). In this case, the orientation of the heading bump changes by 180°, and thus the fly's behavior at an arena heading θ is determined by the drift rate $\nu(\theta + 180^\circ)$ at the phase-shifted heading $\theta + 180^\circ$ (middle panel, middle column). The point-wise product between these curves (lower panel, middle column) gives the second contribution to the observed drift rate. Finally, the sum of these two point-wise products produces the observed drift rate as a function of arena heading (lower panel, right column). The observed fixation duration is proportion to the inverse of the observed drift rate (upper panel, right column). **d)** Same as (**c**), but illustrated for the observed saccade bias, denoted by the probability of clockwise saccades $P_{CW}(\theta)$.

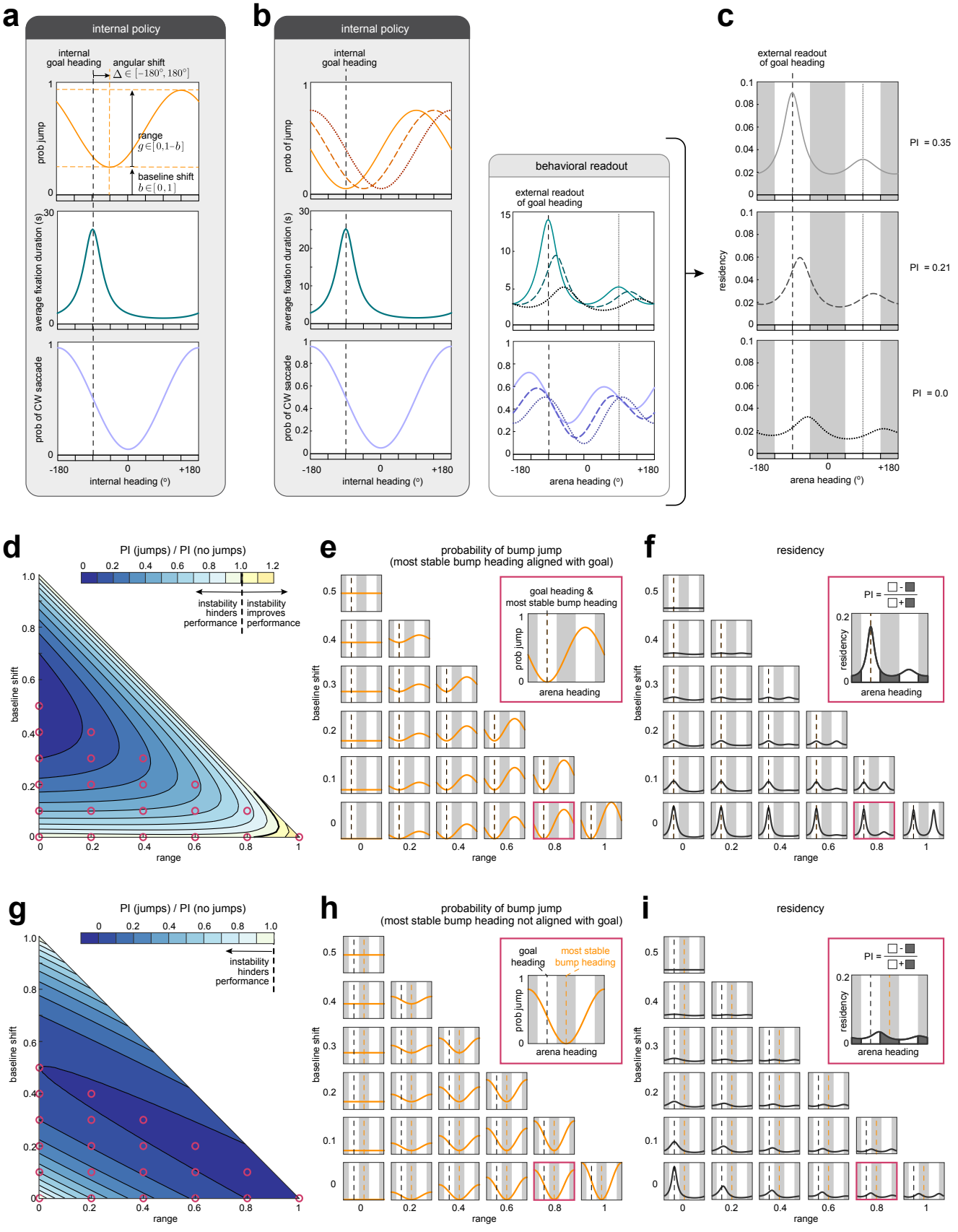

Figure S5: Coherent behavior depends on the statistics of bump jumps. a) The model fly's behavioral policy is specified by a drift rate (which controls the duration of fixations; green curve) and a saccade bias (purple curve) that are both structured as a function of internal bump heading relative to a internal goal heading (dashed vertical line). We assume that this internal bump heading is unstable, and that the heading instability is specified by the probability that the bump will jump from a given bump heading (orange curve). The degree of structure in the instability is controlled by three parameters: the angular shift Δ between the most stable heading (lowest probability of a jump) and the goal heading, the range g, and the baseline shift b. **b)** Varying the angular shift of the heading instability relative to the goal heading (upper panel, left column) leads to a behavioral readout (right column) that is shifted with respect to the goal heading, and whose structure is degraded with respect to the case in which the angular shift is 0° (compare the dashed and dotted green curves, corresponding to angular shifts of 45° and 90°, respectively, to the solid green curve, corresponding to an angular shift of 0°). Note that in all three cases, the internal policy governing fixations and saccades (middle and bottom panels, left column) is kept fixed. **c)** The behavioral readout of fixation duration and saccade bias can be used to compute the expected residency at different locations in the arena, and —in the presence of punishment— the expected PI score (see *Modeling Methods: Modeling the statistics of bump jumps*). Here, we consider the punishment structure used in this assay (with danger zones indicated by the gray bars), and we consider a scenario in which the internal goal heading is aligned with the center of the left-hand safe zone. **d-i)** Expected PI scores for the scenario in which the most stable heading is aligned with the goal heading (**d-f**) and the scenario in which the most stable heading is shifted by 90° with respect to the goal heading (**g-i**). **d)** Expected PI scores as a function of the range and baseline shift of the heading instability, for an angular shift $\Delta = 0^\circ$. PI scores are scaled relative to the scenario in which $g = b = 0$ (i.e., the bump never jumps). Blue and yellow colors indicate parameter combinations for which the heading instability degrades versus improves performance, respectively. **e-f)** Probability of bump jumps (**e**) and residencies (**f**) for the parameter combinations marked by red circles in (**d**). **g-i)** Same as (**d-f**), but for the case in which the most stable heading is shifted by 90° with respect to the goal heading (i.e., $\Delta = 90^\circ$).

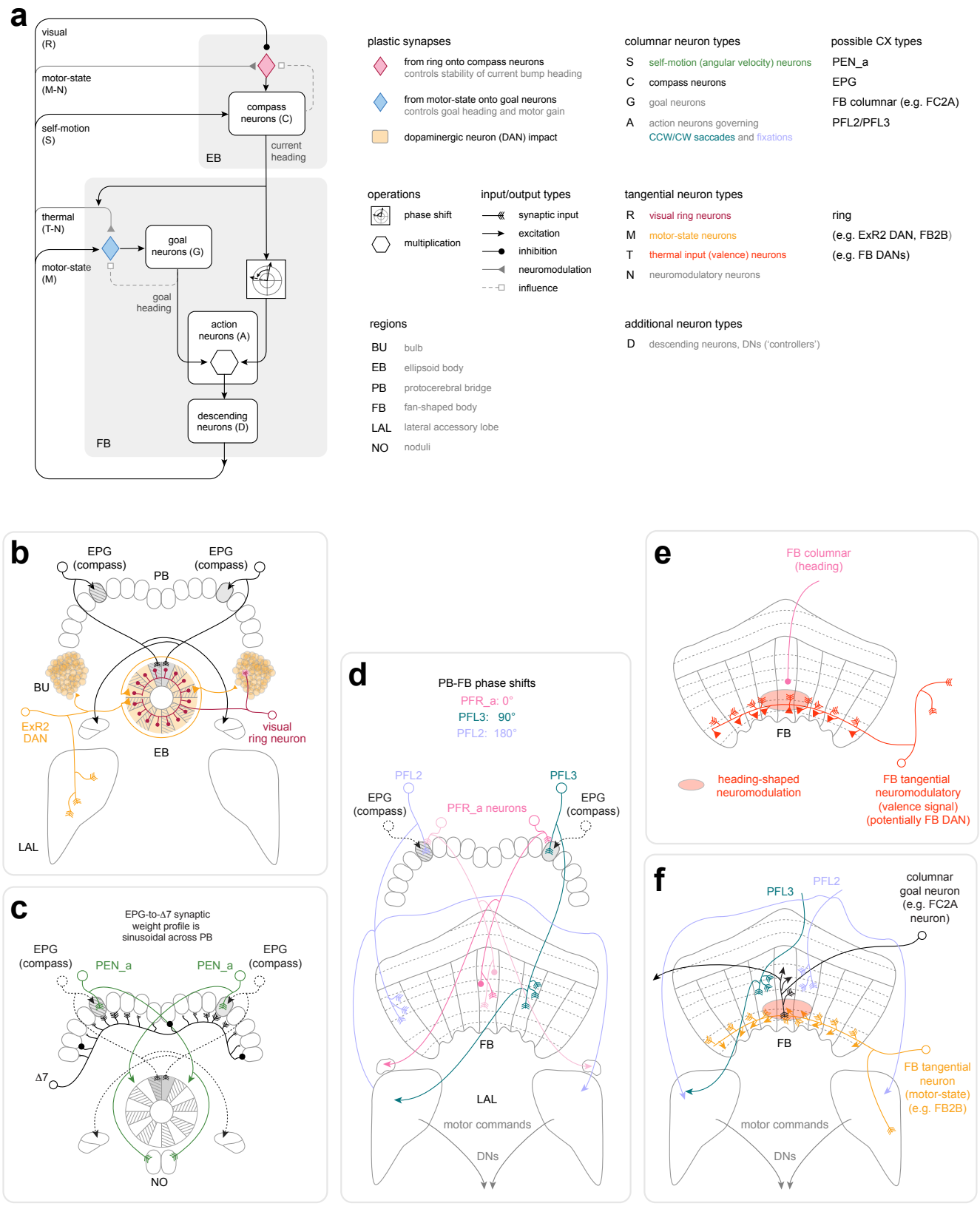

Figure S6: Support for the circuit model from central complex anatomy. **a)** The circuit model drives behavior based on the overlap between the current and goal headings, carried out by multiplicatively combining goal neuron activity with phase-shifted compass neuron activity. Plasticity in weights onto compass neurons (red diamond) and goal neurons (blue diamond) enables the co-evolution of internal mappings of sensory surroundings and of goals within those surroundings. As illustrated in the remaining panels, this circuit model relies on abstracted elements and circuit motifs of the fly CX; see *SI: Linking the Conceptual Model to Known Anatomy* for more details. Note that although the compass neurons are thought to carry HD information, the constraints of our tethered flight setup made heading equivalent to HD. *SI: Linking the Conceptual Model to Known Anatomy* also discusses how the model might also be relevant for, and operate on, true heading (traveling direction). **b)** Flexible mapping of a visual scene onto the fly's HD representation. The fly's HD representation tethers to sensory cues in the environment, which are conveyed to the ellipsoid body (EB) through multiple classes of ring neurons. Visual ring neurons (red) receive inputs in the bulb (BU) and project to the the EB, where they make all-to-all inhibitory connections onto, and receive feedback from, compass (EPG) neurons. Plasticity between ring and compass neurons ensure a self-consistent mapping between the sensory world and the internal HD representation. ExR2 dopaminergic neurons (DANs, orange) receive input in the lateral accessory lobe (LAL) and project to the EB and both BUs, making these DANs a likely source of motor-state-dependent neuromodulation in the EB. This neuromodulation could drive plasticity in the mapping between ring and compass neurons. **c)** Maintaining, updating, and formatting the HD bump. In addition to being tethered to visual and other sensory input, the location of the HD bump in the EB is also determined by multiple classes of columnar neurons (PB-EB neurons) that link the protocerebral bridge (PB) and the EB. Most notably, PEN_a neurons tuned to the HD and the fly's angular velocity update the HD bump's position based on self-motion input. Note that their projection pattern from the PB to the EB ensures that their input is phase-shifted relative to their compass neuron inputs. The sinusoidal weight profile of compass neurons onto downstream $\Delta 7$ neurons, which inhibit other compass neurons, ensures that the HD bump maintains a sinusoidal shape. **d)** PB-FB columnar neurons that link the PB and the fan-shaped body (FB) have a range of phase shifts. The PFR_a neurons are shown as an example of a PB-FB columnar neuron type that has a 0° phase shift between its PB and FB projections, that is, a bump inherited from the compass neurons in the PB would be transferred to the matching column of the FB. This is in contrast to the PFL2 neurons, which project to FB columns that are 180° shifted in phase relative to their PB arbors, and to the PFL3 neurons, whose projection patterns produce a 90° contralateral phase shift. These phase shifts provide the basis for the prestructured behavioral policy in the model. Note also that PFL2 neurons project to the LAL on both sides of the brain, thereby making them suitable for fixation control, and that the PFL3 neurons project unilaterally, making them ideal saccade controllers. The LAL is innervated by multiple classes of descending neurons (DNs), which project to motor centers in the ventral nerve cord. **e)** Valence signals shaped by an HD or heading bump. The FB is innervated by numerous tangential neurons that project throughout specific FB layers, where they show both presynaptic and postsynaptic specializations. Specifically, many FB tangential neurons receive columnar inputs from neurons that could carry HD or heading bumps without the phase shifts that characterize PFL FB activity. Some FB tangential neurons are known to be neuromodulatory; for example, FB DANs like FB2A, FB4L, and FB4M, and may carry valence signals, such as those associated with heating or cooling. Their neuromodulatory signals in the FB are likely to be shaped by their local columnar input, motivating our assumption of heading-shaped neuromodulation as driving plasticity in the goal weights. **f)** A circuit motif for updating goal weights. In our model, FB tangential neurons carrying a heading-shaped valence signal drive plasticity at synapses between motor-state-dependent tangential neurons and largely intrinsic FB columnar neurons. Potential candidates for such motor-state neurons include FB2B neurons that receive input in the LAL and other areas with motor signals. These and other candidate FB tangential neurons receive input from FB DANs and other neuromodulatory FB tangential neurons, and themselves make synapses onto multiple classes of FB columnar neurons that could function as the goal neurons we assume in our model. Reinforcement would modify the strengths of these synapses, leading to different profiles of goal-weighted activity in the columnar neurons when the fly is active. Here we show one possible FB columnar neuron type, FC2A, which happens to receive input from tangential neurons in ventral layers of the FB and makes synapses onto the putative action neurons, PFL2 and PFL3. In our model, this circuit motif would allow individual PFL neurons to multiplicatively combine information from the goal neurons with phase-shifted versions of the HD bump, thereby enabling the population of PFL neurons to implement a prestructured behavioral policy for turning and fixation relative to the goal.

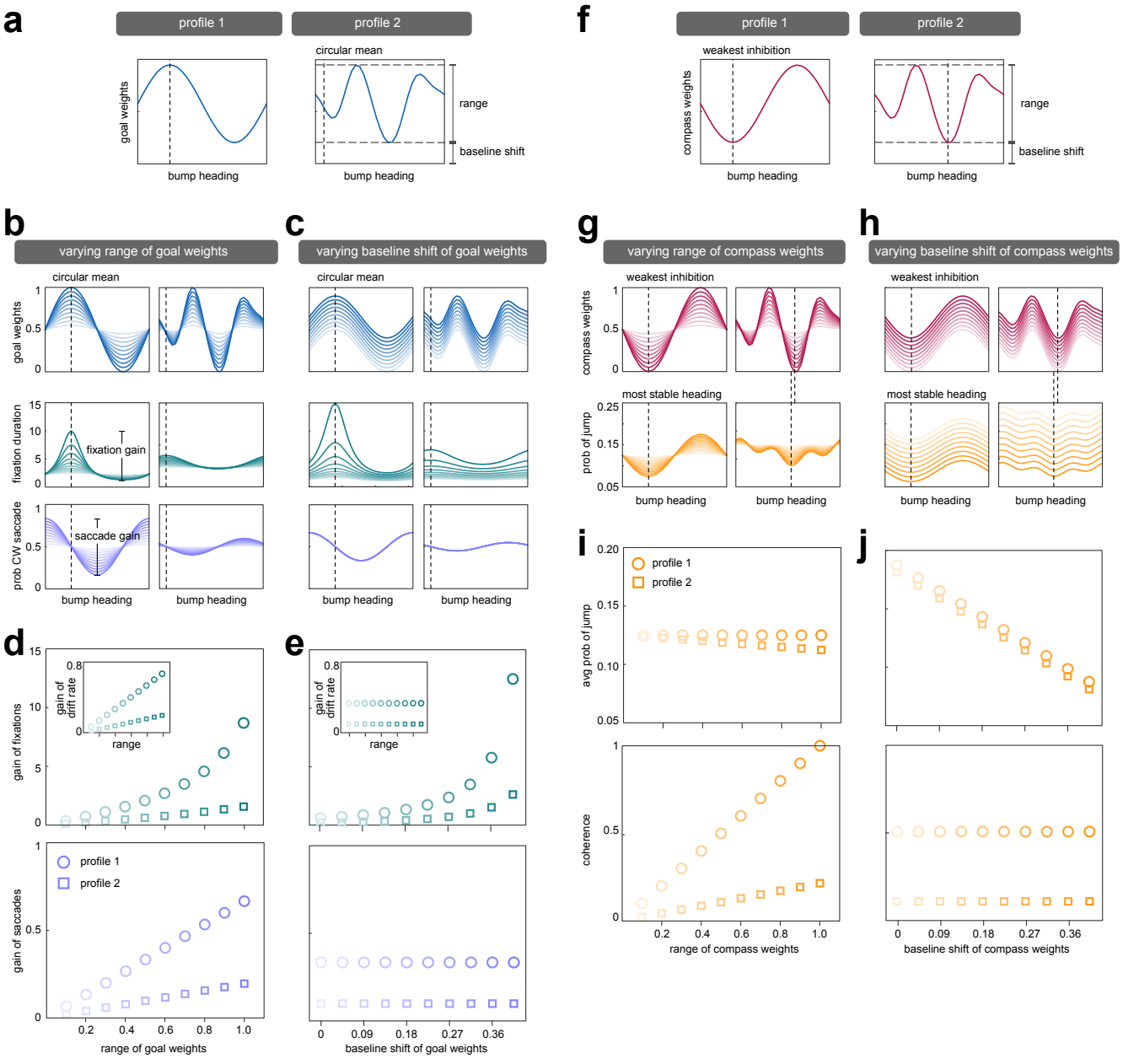

Figure S7: Compass and goal weights control behavioral gain and statistics of bump jumps. a) The profile of goal weights controls the behavioral policy. Here, we use two different weight profiles to illustrate their impact on behavior. Each profile can be described in terms of its baseline shift, range, and circular mean. In panels (**b,d**), we vary the range of the weight profiles while fixing their average (equal to the baseline shift plus half of the range). In panels (**c,e**), we vary the baseline shift of the weight profiles while fixing their range. **b-c)** The circular mean of the goal weights (dashed vertical lines, upper row) determines the location at which fixations are longest (middle row) and to which saccades will drive the HD bump (bottom row), and thus specifies the goal heading. Different weight profiles generate more (left column in (**b,c**)) or less (right column in (**b,c**)) structure in the behavioral output. **d)** The larger the range of the goal weights (darker colors in (**b**)), the larger the gain of fixations (upper panel) and saccades (lower panel). The more singly-peaked the weight profile (circular markers), the greater the impact the range of the weights has on fixations and saccades. Note that the gain of both the drift rate (inset, upper panel) and the saccade bias scale linearly with the range of the goal weights (this can be understood from Eqs.35-36); the duration of fixations, which is inversely related to the drift rate, scales supralinearly with the range of the goal weights. **e)** Increasing the baseline shift of the goal weights (darker colors in (**c**)) increases the gain of fixations, but does not impact the gain of saccades (this can again be understood from Eqs. 35-36). As in (**d**), the more singly-peaked the weight profile (circular markers), the larger the overall gain of fixations and saccades, and the greater the impact the range of the weights has on the gain of fixations. **f)** The profile of compass weights captures the net inhibition from ring neurons onto compass neurons, and controls the stability of the bump heading. As in (**a**), we use two different weight profiles to illustrate their impact on bump stability, and we describe each profile in terms of its baseline shift, range, and angular location of weakest inhibition. **g-h)** The probability that the heading bump will jump depends on the difference in inhibition between headings separated by 180°. As a result, the most stable heading with the lowest probability of a jump (dashed line, second row) can differ from the location of weakest inhibition (dashed line, first row). **i)** The larger the range of the compass weights (darker colors in (**g**)), the less likely the HD bump is to jump (upper panel) but the more coherent the jumps (lower panel; coherence is measured as the fraction of jumps made toward, rather than away from, the most stable bump heading; see *Modeling Methods: Compass weights control the probability and coherence of bump jumps*). The more singly-peaked the weight profile, the greater the impact the range of the weights has on the coherence of bump jumps. **j)** Increasing the baseline shift of the compass weights (darker colors in (**h**)) increases the net inhibition onto compass neurons and decreases the overall probability that the bump will jump (upper panel), but does not affect the coherence of bump jumps. As in (**i**), the more singly-peaked the weight profile (circular markers), the higher the coherence of bump jumps.

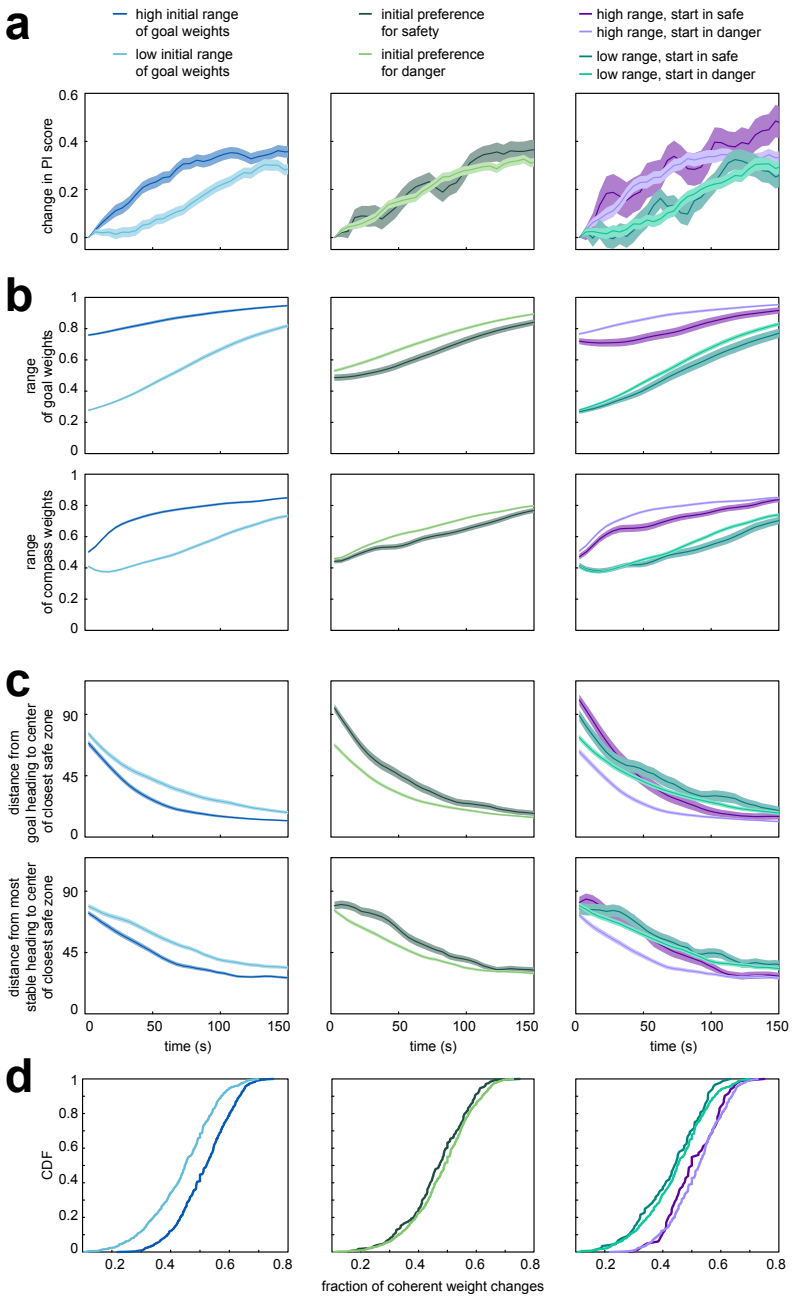

Figure S8: Structure in behavior impacts learning. Performance measures as a function of time (**a-c**) and accumulated across time (**d**) for 1000 model simulations, separated by the initial degree of structure in the behavior (left column; initial range of goal weights above or below a threshold of 0.5, respectively), initial preference for safety versus danger (middle column; initial distance to safety above or below a threshold of $45°$, respectively), and both behavioral structure and initial preference (right column). Shaded regions: mean+/- standard error. **a)** The initial range of goal weights, which controls the degree of structure in the behavior (see SI Fig S7), impacts the subsequent changes in PI scores to a much greater extent (left, right panels) than does initial preference for safety versus danger (middle, right panels). **b-c)** A higher initial range of goal weights leads to a faster increase in the range of both goal and compass weights (**b**), and a faster shift of both sets of weights toward safety (**c**). **d)** Because a high initial range of goal weights leads the compass weights to stabilize more quickly (see lower left panels in (**b,c**)), the heading bump will be more likely to jump coherently toward to the goal heading. This will lead to a larger fraction of coherent weight updates that shift the goal heading toward its final location. In contrast, a low initial range of goal weights will lead to a slower stabilization of compass weights, more incoherent bump jumps, and ultimately more incoherent weight updates that shift the goal heading in opposing directions with respect to its final location.

METHODS

Experimental Methods

Fly culture. Parental flies were grown sparsely on Wurzburg food in bottles for at least 6 generations [97]. Crosses were first done in vials then transferred to bottles after 1-3 days, followed by transferring to a new bottle every day to limit F1 density. 10 males and 25 virgins were used for each cross. The day after eclosion, F1s were transferred to a new bottle with a piece of kimwipe for self-cleaning and transferred again to a new bottle with kimwipe the day before imaging or behavioral experiments. All experiments were performed with 5-6 days old female flies.

Visual arena. A blue LED circular arena [38] was assembled with 44 panels (4 rows and 11 columns, spanning 120° in elevation and 330° in azimuth), with the LED emission peaking at 464 nm (Bright LED Electronics Corp., BM-10B88MD). Two layers of blue filter (Roscolux #59) were laid on top of the LED panels to allow 0.04% transmission. Each fly was tethered at the end of a tungsten wire and positioned in the center of the arena. An 880 nm LED (Digi-Key, PDI-E803-ND) illuminated the fly from above. A custom-built wingbeat analyzer (University of Chicago Electronics Shop) measured the wingbeat frequency and amplitudes for both wings. Yaw turning was computed as the left minus right wingbeat amplitude. A computer (Dell, R5500) controlled the timing of the experiments through a data acquisition device (National Instruments, USB-6229 BNC) and sampled the flight parameters at 1 kHz. This is in contrast to most classical visual learning studies, which have relied on torque meters to measure the fly's instantaneous torque and drive the rotation of a paper drum imprinted with visual patterns [18, 39].

Flight visual learning. The 360° yaw space around the fly was divided into 4 quadrants. A single horizontal blue bar (37.5°w x 11.25°h) was displayed in each quadrant, with alternating elevations at +/- 30°, such that the pattern repeats every 180°. Throughout the assay, the fly had closed-loop control of the visual pattern it was flying towards. The unconditioned stimulus (US) as punishment was a fiber-coupled infrared laser (Edmund Optics, 975 nm, 400 mW) modulated by a 10 kHz square wave with varying duty cycles output from a function generator (Agilent, 33220A, 20 MHz) and gated by the specific positions of the arena pattern such that either the higher bar quadrant or the lower bar quadrant was accompanied by the laser punishment aimed at the back of the fly. The laser was turned off during the pre-training naïve trials and post-training memory/probe trials. The visual pattern was jump-rotated randomly to a new position after every trial. A 100 ms air puff towards the fly was triggered whenever the fly stopped flying. However, only data during flight from flies that flew continuously without stop or puffing for more than 60 s in all 3 trial types were included for further analysis. All visual stimulation and behavior parameters were recorded with a data acquisition device (National Instruments, USB-6229 BNC). During no-laser mock experiments, the US laser was not turned on.

Fly preparation for imaging during flight. Flies were transferred to a polypropylene tube using a custom 3D-printed funnel positioned on the top of the opened bottle, then anaesthetized in a custom brass cold plate at 4°C. The largest female was selected to fit onto a custom aluminum mounting bridge cooled to 4°C, and held down with vacuum suction ventrally. The bridge was then rotated to hold the fly upside down and an inverted custom laser-milled PEEK holder pushed up the fly's head from below, with a center hole lined up under the head. Small drops of UV-activated epoxy were used to glue the fly head, thorax and the back of the head capsule to the holder. Another small drop was used to glue the proboscis. The eyes were kept completely below the holder to allow unhindered visual stimulation and most of the back head plate was exposed through the center hole in the holder. The legs were left intact and the wings kept free to flap during flight because of the inverted pyramid shape of the holder. The back plane of the head was angled at approximately 26° to match the angle of the visual arena under the two-photon microscope. For imaging experiments, we used an LED arena with 18 panels (3 rows and 6 columns, spanning 90° in elevation and 180° in azimuth). For flight experiments, only flies that could fly continuously for 90 s while maintaining closed-loop stripe fixation after mounting were selected. Artificial hemolymph as described previously [98] was used to fill the holder reservoir from the top. A window was carefully opened on the back head capsule with a tungsten dissection probe and fine forceps and the trachea underneath were gently picked away to allow optic access to the brain [99].

Two-photon calcium imaging. Calcium imaging was performed on a two-photon microscope (Bruker Nano, formerly Prairie Technologies). A Chameleon Vision II or Discovery laser (Coherent) tuned to 920 nm was used with the power adjusted to the lowest sufficient level, usually between 3 and 20 mW at the sample. A resonant

galvanometer mirror was used to scan the laser beam along the x-axis at 8 kHz, resulting in a frame rate of 60 Hz with 256 by 256 resolution. For volume imaging, a piezo motor drove the 40x objective (Olympus, LUMPlanFl/IR, NA 0.8) along the z-axis. The 2-plane z-stack acquisition was repeated over time throughout the trial at a rate of 14.5 volumes/s. The green and red channel signals, when applicable, were collected through a set of dichroic mirror (575 nm) and band-pass filters (525 + 35 nm for green, 607 + 22.5 nm for red). A GaAsP photomultiplier tube (Hamamatsu, 7422PA-40) was used to acquire data from each channel. Each imaging series was triggered from the experiment-controlling computer.

Data Analysis

All data analysis was performed in MATLAB (MathWorks Inc., Natick, MA).

Partitioning behavior into fixations and saccades. All data analyses were performed after segmenting behavioral traces into fixations and saccades. We developed a custom algorithm to perform this segmentation (SI Fig S2a). We first filtered the difference in wingbeat amplitude between left and right wings, A^{WB}, using a bandpass filter of order 6, with a lower cutoff frequency of 0.1 and an upper cutoff frequency of 10 (we'll denote this filtered signal as $\tilde{A}^{\mathrm{WB}}$). We then used sign changes in the filtered amplitude to segment the trajectory into a set of individual turns; each turn in this set was thus defined as a sequence of time points $\{t\}$ for which $\tilde{A}^{\mathrm{WB}}_{\{t\}}$ had a consistent sign.

A turn that produces a sustained nonzero difference in wingbeat amplitude $\tilde{A}^{\mathrm{WB}}$ will lead to changes in the arena heading x in the opposite direction. We used this to define a quantity $s_t = -\tilde{A}^{\mathrm{WB}}_t|\Delta x_t|$, where $\Delta x_t = x_{t+1} - x_t$ is the instantaneous change in arena heading (allowing for wrapping between pixel 96 and pixel 1). This quantity measures the coherence between differences in wingbeat amplitude and changes in arena heading; s_t will be large in magnitude during times when changes in wingbeat amplitude lead to large and coherent changes in arena heading, and will be zero when changes in wingbeat amplitude do not lead to a change in arena heading. We thus used this signal to select turns that fall into the former category, where s is large in magnitude.

To this end, we first selected candidate saccades as those turns that led to a total change in arena heading of at least 2 pixels ($7.5°$). For this subset of turns, we used s_t to refine the beginning and end of individual turns. We defined the beginning of the turn as the timepoint t_{start} for which there was the largest instantaneous change $\Delta s_t = s_{t+1} - s_t$, and the end of the turn as the first timepoint thereafter for which s_t dropped below $1/4$ of its maximum value (i.e., $t_{\mathrm{end}} : s_t < \frac{1}{4} s_{t_{\mathrm{start}}}$). The remaining timepoints ($t < t_{\mathrm{start}}$, and $t > t_{\mathrm{end}}$) were segmented as separate turns. We repeated this process until all large turns had been refined in this way.

This resulted in a refined set of turns; these turns included both the candidate saccades that led to a change in arena heading, and the small turns that did not lead to a change in arena heading. We removed all turns during which the wingbeat frequency f^{WB} dropped below a threshold of $f^{\mathrm{WB}}_{\mathrm{min}} = 1$. We then ranked each remaining turn according to a quantity $r(\mathrm{turn}) = |\langle s_{t_{\mathrm{start}}:t_{\mathrm{end}}} \rangle (x_{t_{\mathrm{end}}} - x_{t_{\mathrm{start}}})|$ that combines the average change in wingbeat amplitude $\langle s_{t_{\mathrm{start}}:t_{\mathrm{end}}} \rangle$ with the total change in arena heading $(x_{t_{\mathrm{end}}} - x_{t_{\mathrm{start}}})$. This quantity will be largest for turns that are large and fast, which comprise a small fraction of the entire set of turns. We thus used an outlier detection procedure to identify those turns for which r exceeded a threshold $r_{\mathrm{thresh}} = Q3 + \exp(3M) * 1.5 * \mathrm{IQR}$. Here, $Q3$ is the third quartile (or 75%) of r, IQR is the inter-quartile range, and M is a skewness estimated using the med-couple of r [100]. r_{thresh} was estimated separately for individual flies and trials, based on the distributions of turns produced by the given fly in the given trial.

The set of turns for which $r > r_{\mathrm{thresh}}$ were classified as saccades, and the periods of time between each saccade were classified as fixations (SI Fig S2a). We described these behavioral modes in terms of their duration and their angular velocity; here, the angular velocity (in degrees/ms) is given by $\omega = -50(360/96)\tilde{A}^{\mathrm{WB}}$, where the factor 50 converts between wingbeat amplitude and pixels/ms, and the factor $(360/96)$ converts from pixels/ms to degrees/ms. Saccades were characterized by short durations and large average angular velocities, while fixations were characterized by long durations and low average angular velocities (SI Fig S2b).

Characterizing fixation properties. Individual fixations varied substantially in their duration, and the distribution of these durations was heavy-tailed. We therefore considered three putative heavy-tailed distributions: log-normal, inverse Gaussian, and generalized Pareto. We fit each of these three distributions to the distribution of fixation durations $P(\Delta t)$ under two different conditions: when fixations were accumulated across flies within a given trial, and separately when fixations were accumulated across trials for a given fly. We performed this fitting for laser-trained and no-laser control flies.

Prior to fitting, we removed fixations whose durations were below a variable threshold $\Delta t_{\mathrm{thresh}}$. We then evaluated fitting performance for 25 evenly spaced values of $\Delta t_{\mathrm{thresh}}$ between 20 and 500 ms. We used the MATLAB function *fitdist.m* to perform the fitting, and we used the Bayesian information criterion (BIC) to evaluate fits. We found that the inverse Gaussian distribution, $\mathrm{IG}(\Delta t;\mu,\lambda)$, was the best-fitting distribution across a majority of scenarios (trials or flies) for thresholds between 100 and 300 ms; within this range, a threshold of 200 ms maximized this number of scenarios for which the inverse Gaussian was the best fit. We therefore performed the remainder of our analysis on fixations whose duration exceeded $\Delta t_{\mathrm{thresh}} = 200$ ms.

The inverse Gaussian distribution is characterized by two parameters: a mean μ and a shape parameter λ. This distribution can be generated by a drift diffusion to bound process with a mean drift rate ν, spread η^2, and bound a (SI Fig S2g). This process yields an inverse Gaussian distribution $P(\Delta t) = \mathrm{IG}(\Delta t; a/\nu, a^2/\eta^2)$ whose parameters $\mu = a/\nu$ and $\lambda = a^2/\eta^2$ are defined in terms of the parameters of the diffusion process. When we compared the best-fitting values of μ_F and λ_F across different datasets (where the subscript F denotes that parameters were fit to the distribution of fixations), we found that the variability in these parameters was consistent with a drift diffusion process with a variable drift rate ν_F but fixed spread η_F^2 and bound a_F. To illustrate this, note that the mean $\mathrm{mean} = \mu$ and variance $\mathrm{var} = \mu^3/\lambda$ of the inverse Gaussian distribution satisfy $\log(\mathrm{var}) = 3\log(\mathrm{mean}) - \lambda$. If the variability in the fit parameters can be explained by changes in ν alone (with fixed η^2 and a), the plot of $\log(\mathrm{var})$ versus $3\log(\mathrm{mean})$ will be well-described by a line of slope 1 and fixed offset $-\lambda = -a^2/\eta^2$. SI Fig S2h shows this comparison when the mean and variance are computed from the fit parameters (filled markers) versus estimated directly from the data (open markers), along with a line of slope of 1 and best-fitting offset $\lambda_F = 0.63$ (dashed line). We found that this provided a better fit than a model in which the bound is variable, and the drift rate and spread are fixed (dotted line).

We used this result to posit that fixations are controlled by a drift diffusion process with a fixed spread η_F and bound a_F, but an adaptive drift rate ν_F. Because there are three parameters of the drift diffusion process but only two parameters needed to define the inverse Gaussian distribution, we are free to choose one of the drift diffusion parameters and fit the other two. We chose to set $\eta_F = 1$, which requires that $a_F = 0.79$ (thus satisfying $\lambda_F = a_F^2/\eta_F^2 = 0.63$). When we restricted our analysis to fixations that were initiated within the danger zone during the first 60s of the first training trial (and remained within the danger zone for 95% of their duration), we found that they were well fit by the same process, but with a higher drift rate and thus shorter average duration (red star in SI Fig S2h). The reduction in fixation duration in response to heat can thus be captured by an additional "reflexive" drift process with drift rate $\nu_F = 0.38$, spread $\eta_F = 1$, and bound $a_F = 0.79$.

Characterizing saccade properties. Individual saccades varied in both their speed and duration. We found that the average duration of saccades depended on their average angular speed; we thus began by characterizing the distribution of angular speeds $P(\omega)$, where ω is computed by averaging the instantaneous difference in wingbeat amplitude over the duration of a saccade. The angular change in heading can be computed from this via $\Delta\theta = \omega\Delta t$. We then characterized the distribution of durations conditioned on speed, $P(\Delta t|\omega)$.

Prior to fitting, we removed saccades whose speeds were below a threshold $\omega_{\mathrm{thresh}} = 0.1$. We used the MATLAB function *allfitdist.m* to fit 16 different parametric distributions to the distribution of speeds $P(\omega)$ under two different conditions: when saccades were accumulated across flies within a given trial, and separately when saccades were accumulated across trials for a given fly. We performed this fitting for both laser-trained and no-laser control flies, and we used BIC to evaluate fits. We found that the lognormal distribution $\mathrm{logn}(\Omega;\varphi,\sigma^2)$, with location φ and scale σ^2, was the best-fitting distribution across the majority of conditions. We then computed the directional bias in saccades, measured as $(N_{\mathrm{CW}} - N_{\mathrm{CCW}})/(N_{\mathrm{CW}} + N_{\mathrm{CCW}})$, where N_{CW} and N_{CCW} respectively denote the total number of clockwise and counter clockwise saccades taken within a single trial. We found that this bias also varied across trials (SI Fig S2d upper), suggesting that flies can adaptively control directional bias of their saccades.

The distribution $P(\omega)$ specifies the probability of initiating saccades of different speeds. For saccades of a given speed ω, there is significant variability in their duration (SI Fig S2f). To characterize this variability, we considered 36 equally spaced values of ω between 0.25 and 2. For each value of ω, we used the MATLAB function *allfitdist.m* to determine the parametric function that best fit the distribution of saccade durations $P(\Delta t|\omega)$ accumulated across flies, trials, and datasets. We found that these distributions were best fit by an inverse Gaussian distribution with fixed spread $\eta_S = 1$ and bound $a_S = 0.57$ but variable drift rate ν_S (SI Fig S2i), analogous to fixations. In this case, the drift rate increased nonlinearly with the speed $|\omega|$ (inset of SI Fig S2i); we used a least-squares fit to determine the parameters of the best-fitting sigmoid $f(|\omega|) = f_M/[1+\exp(-k(|\omega| - \omega_0))] - f_0$; these where given by $f_M = 2.6$, $k = 7.36$, $\omega_0 = 0.32$, and $f_0 = -1.01$. Thus, for a saccade initiated with speed ω, the duration can be generated via a drift diffusion process with a velocity-dependent drift rate $\nu_S = f(|\omega|)$, and fixed values of $\eta_S = 1$ and $a_S = 0.57$.

Inferring the structure of a behavioral policy. Together, the analysis of fixations and saccades enable us to construct a behavioral policy that accounts for variability in the initiation, speed, and duration of both fixations and saccades (SI Fig S2j). Each behavioral mode (fixation versus saccade) is generated via a sequence of three steps: (i) Select an angular velocity by sampling from a parametrized distribution. For fixations, we approximate the angular velocity to be zero. For saccades, we sample the magnitude of the angular velocity from a lognormal distribution, and we take the directional bias (corresponding to the likelihood of initiating a clockwise versus counter clockwise saccade, which specifies the sign of the angular velocity) to be an adaptive parameter. (ii) Generate the duration via an online drift diffusion process with a variable drift rate. For fixations, we take this drift to be an adaptive parameter. For saccades, this drift is determined by the angular velocity selected in step (i). (iii) Determine the resulting change in heading, which is proportional to the product between the average angular velocity and the duration.

In the main text, we considered a simplification of this full behavioral policy in which the angular size of saccades (measured in deg) was directly sampled from a lognormal distribution with parameters $\varphi_S = 3.89$ and $\sigma_S = 0.54$ (fit to the distribution of saccade sizes across all flies, trials, and datasets; this generates saccades with an median angular size of $49°$), and assuming a fixed saccade duration of $t_S = 320$ ms (equal to the median saccade duration measured across all flies, trials, and datasets). Table 1 summarizes these choices.

Data selection. For all analyses in the paper, we selected those fixations that exceeded a duration of $\Delta t_{\text{thresh}} = 200$ms, and those saccades that exceeded an average angular velocity of $\omega_{\text{thresh}} = 0.1$/ms. For a given fly on a given trial, this thresholding resulted in a subset of selected time points; we kept only those trials for which this set of selected time points exceeded a total of 70s.

Residency. Figs 1b and 5g show residency in different regions of the arena. In Fig 1b, we computed this residency with respect to individual pixels in the arena as a function of their location relative to the safe and danger zones. Due to the two-fold symmetry of the visual scene, this resulted in pairs of pixels that shared the same location relative to safety/danger. For each fly, we summed the residency (measured in ms) for each such pair of pixels, and we divided this by the total number of timepoints that met the selection criteria described above. Fig 1b shows the average and standard error of this residency, measured across flies, for each such location in the arena. The dashed horizontal line shows the residency that would be expected if flies spent equal amounts of time at all locations within the arena, given by $1/48 = 0.0208$. In 5g, we computed the residency with respect to individual pixels in the arena for two individual flies, without combining pairs of pixels that shared the same location relative to safety/danger.

Performance scores. Fig 1c shows the performance index (PI) scores, averaged across flies on each trial. PI scores were computed as $\text{PI} = (T_{\text{safe}} - T_{\text{danger}})/(T_{\text{safe}} + T_{\text{danger}})$, where T_{safe} and T_{danger} denote the total time spent in safe and danger zones, respectively [18].

Aligning to individual preference. Fig 2h shows the average duration of fixations and directionality of saccades after aligning the behavioral data to the arena preference of each fly within each trial. To perform this alignment, we first computed the fraction of time that the fly resided at each of 96 orientations (corresponding to 96 pixel locations) within the arena. We then constructed an idealized residency profile that took a peak value of one at a central set of two pixels, and decayed linearly to zero over a span of 24 pixels in either direction (CW and CCW). We shifted this idealized profile with respect to the true residency profile of the fly, and we identified the preferred orientation as the one that maximized the overlap between the idealized and true residency profiles.

Computing heading-dependent averages. To perform the heading-dependent averages shown in Fig 2h, we first selected the set of saccades and fixations taken by each fly within probe trials 2, 5, 8 and 9. We binned saccades according to the arena heading at which they were initiated; we binned fixations according to the average heading computed across the duration of the fixation. We used 16 evenly spaced bins, each spanning $22.5°$. We then computed the average direction of saccades initiated within each bin, and similarly the average duration of fixations within each bin. Fig 2h shows the average and standard error of these quantities, measured across flies.

Computing calcium transients. For volume imaging of EPG GCaMP activity (Fig 3e-f), we used two z-planes that together captured the dorsal and ventral halves of the EB. The image stack at each time step was converted into a summed intensity projection that was used for further analysis. We manually divided the EB into 32 wedge-shaped ROIs to capture population EPG activity in the structure. An additional ROI without any EPG arborization and

outside the EB was selected to estimate background signal, including from leaked LED arena light. Time series of GCaMP activity for all EB ROIs were obtained by taking the average of the fluorescence signal within each ROI at each time step. The calcium transient for each ROI, $\Delta F/F_0$, was computed by subtracting fluorescence in the background ROI from all other ROIs, and using the lowest 10th percentile of background-subtracted fluorescence from each ROI as F0. The resulting time series were filtered using a simple boxcar (moving mean) filter (width 344 ms).

Rather than compute the population vector average (PVA), as in past work [19, 29, 43, 58, 61], we focused here on tracking peaks in EPG population activity ('bump position' or 'bump heading') at each time step. This allowed us to to track offsets between the EB location of the bump and the position of the visual scene at every time point, and to easily visualize changes in offsets, bump jumps, as seen in Fig 3f. Considering the symmetry of the visual scene, we tracked the position of the visual scene using two 180°-offset time series, with the first being shifted by the first offset and the second by a second offset (if present, see below).

Clustering bump offsets. Fig 3g-h shows the offsets between the bump and the visual scene for individual flies. To cluster these offsets, we used the MATLAB function *kde.m* (with 256 mesh points) to perform kernel density estimation of the distribution of offsets for each fly on each trial. We then used the MATLAB function *findpeaks.m* to determine the peaks in this density; we used offset values corresponding to these peaks as our candidate offset values. We then used the same function to determine the minima in this density (using a peak threshold of 10^{-4}), and we used the offset values corresponding to these minima as the bounds between different clusters. We the computed the sum of the density function within these bounds, divided by the sum of the density function over all time, and used this as the fraction of time spent at each offset. Fig 3g-h shows the fraction of time spent at different offsets for individual flies on individual trials.

Characterizing the number and angular separation of offsets. Fig 3i shows the total number of and angular separation between offsets. To construct these histograms, we first computed the fraction of time that the HD bump spent at different offsets relative to the visual scene for each fly on each trial, as described above. We then computed the number of instances (aggregated over flies and trials) that we observed a given number of distinct offsets; these results are shown in the upper panel Fig 3i. For flies that exhibited two or more offsets on a given trial, we computed the angular distance between the dominant two offsets; this histogram is shown in the lower panel of Fig 3i.

Computing HD tuning curves. Fig 3j shows the tuning of EB wedges to different arena headings. We first determined all times (aggregated across all 9 trials) that the visual scene was oriented at a particular angle relative to the fly, and then computed the average fluorescence transients $\Delta F/F$ of each wedge for each given scene orientation (see *Analysis Methods: Computing calcium transients*). The HD tuning curves on the right hand side of Fig 3j show the average tuning of individual wedges as a function of the fly's heading in the arena, i.e., the "arena heading", which differs from the scene orientation by a sign flip.

Determining the locations of bump jumps with the EB. Fig 3k-l shows bump jumps as a function of their location within the EB. To determine the location of the bump jumps, we first determined the relationship between the arena heading and bump phase that minimized the angular distance between successive time points; an example of this relationship is shown in Fig 3k. We took advantage of our previous results (shown in the lower panel of Fig 3i) to select those changes in bump phase between 135° and 225°; this range captured the majority of the bump jumps in our data. For each jump, we marked the location within the EB at which the jump was initiated. We then computed the angular distance from this location to the location of the putative preferred bump heading within the EB. To determine this location, we used the behavioral data for the same fly on the same trial to infer a preferred arena heading, as described above. For a given preferred arena heading, we determined the corresponding location in the EB at which the heading bump spent the most time. We used this as the location of the putative preferred bump heading in the EB. The upper panel of Fig 3l shows the number of jumps that were initiated at a given angular distance from this goal location, divided by the total number of times that the bump visited locations of the same angular distance, for a single fly (accumulated across trials). The lower panel of Fig 3l shows the same histogram, now accumulated across flies. To approximate these jump statistics, we determined the parameters of the best-fitting cosine function that minimized the mean-squared error between the measured and fit values of the histogram.

Note that the behavioral experiments were performed in arenas with a 330° angular span, but the imaging experiments were performed in arenas with a 180° span in the azimuth. Although we cannot rule out the possibility

that the reduced horizontal span of the visual scene in imaging experiments affected the probability of the EPG bump jumping, similar bump instabilities have been reported in both flying and walking flies in symmetric visual settings in larger arenas as well [19, 29, 30].

Measuring the degree of structure in the behavior. Fig 5h-i shows the degree of structure in the behavior of two individual flies and averaged across flies. To compute this, we first aligned the behavioral data to the preferred arena heading for individual flies on individual trials, as described in *Analysis Methods: Aligning to individual preference*. We then computed the degree of structure in behavior with respect to the preferred arena heading by computing the overlap of the arena residency (measured as the fraction of total time spent at each heading) with the sawtooth profile shown in the legend of Fig 5h, specified over the range $[0, 1]$. This overlap measures how coherently the behavior was structured within the quadrants corresponding to the preferred and anti-preferred headings, and penalizes residency in the other two quadrants. In Fig 5h, we measured this gain for two individual flies, and averaged across $n = 44$ flies. In Fig 5i, we partitioned flies into two groups whose overlap was greater than ("more structured") or less than ("less structured") a fixed value of 0.25. We computed the average and standard error in PI scores within each of these two groups.

Measuring distance to safety. Fig 5h-i shows the distance to safety for two individual flies and averaged across flies. To compute this, we first aligned the behavioral data to the preferred arena heading for individual flies on individual trials, as described in *Analysis Methods: Aligning to individual preference*. We then computed the minimum angular distance between this preferred heading and the center of the closest safe zone. In Fig 5h, we measured this gain for two individual flies, and averaged across $n = 44$ flies. In Fig 5i, we partitioned flies into two groups whose minimum distance to safety was either less than ("initial preference for safe") or greater than ("initial preference for danger") half the angular size of the safe zone (i.e., $90°$). We computed the average and standard error in PI scores within each of these two groups.

Modeling

Determining the optimal policy for maintaining a preference. Fig 2f-g shows the average duration of fixations and directionality of saccades that result from training a flexible RL agent to either exhibit a preference for a specific visual pattern (Fig 2f) or exhibit a preference for an arena heading (Fig 2g). The learning algorithm is described in detail in *SI: Reinforcement learning framework* (see Algorithm S8); the parameters used in the model are summarized in Table 1.

Briefly, in both cases, we learned a single set of weights $\vec{\omega} = \{\vec{\omega}_F, \vec{\omega}_S\}$ that specify the duration of fixations ($\vec{\omega}_F$) and directionality of saccades ($\vec{\omega}_S$) as a function of angular orientation (via a set of 16 von Mises radial basis functions), and we reported the resulting behavior when averaged over 100 different training runs. In the first case (shown in Fig 2f), the weights specify behavioral properties relative to the orientation of visual patterns; thus, whenever the model fly is orienting toward either high bar, the same set of weights will be used to specify the behavioral output. By comparison, in the second case, the weights specify behavioral properties relative to the arena heading, regardless of the specific visual pattern shown at that heading. In both cases, reinforcement was delivered as a function of the current orientation of the pattern/heading relative to the preferred orientation of the pattern/heading; we assumed this reinforcement decayed linearly away from the preferred pattern/heading (see *SI: Reinforcement learning framework* for more details).

Prior to each training run, we initialized the set of flexible policy parameters $\vec{\omega}_F = 0.1$ and $\vec{\omega}_S = 0$, and we randomly initialized the arena heading to one of 96 evenly-spaced values between 0 and $360°$. Following each training run, we used Eq. 17 to evaluate the drift rate of fixations, $\nu(\theta; \vec{\omega})$, and the probability of rightward saccades, $p_R(\theta; \vec{\omega})$, as a function of bump heading θ given the learned parameters $\vec{\omega}$. We then computed the average duration of fixations $a_F/\nu(\theta; \vec{\omega})$, and the average directionality of saccades $2p_R(\theta; \vec{\omega}) - 1$. We averaged these across training runs to produce the curves shown in Fig 2f-g.

Coupling the policy to an unstable heading. Fig 2i illustrates the optimal policy derived in Fig 2g (and described above) coupled to an unstable internal representation of heading. To construct the heading instability, we assumed that the internal heading could jump between orientations that correspond to symmetric views of the visual scene; for the two-fold symmetric scene used here, this corresponds to a jump of $180°$. We further assumed that the jumps occurred probabilistically, and were least likely to occur at the preferred heading and most likely to occur at the symmetric (or "anti-preferred") heading. We used a cosine function to parametrize this probability; this is given in Eq. 21, with parameter values $G_J = 0.7$, $\theta_G = -90°$, and $B_J = 0.05$.

Illustrating the impact of a structured heading instability. To illustrate the impact of an unstable bump heading on the behavioral readout of fixation durations and saccade directions, as shown in SI Fig S4, we used Eq. 23 with parameter values given in Table 1.

Modeling the statistics of bump jumps. SI Fig S5 shows how a structured heading instability impacts PI scores. As described above in *Modeling Methods: Illustrating the impact of a structured heading instability*, we used Eq. 23 with parameter values given in Table 1 to illustrate the fixed policy shown in SI Fig S5a-b. In SI Fig S5b, we varied the angular shift of the curve governing the probability of bump jumps, using shifts of $0°$, $45°$, and $90°$. In SI Fig S5c, we used the external readout of fixation duration and saccade probability to compute the expected residency at different headings, assuming long timescales for behavioral sampling. To do this, we used the probability of saccades, together with the distribution of saccade sizes, to compute the expected number of visits to each heading. We then used this distribution to weight the duration of fixations at each heading, resulting in the expected fraction of time spent at each heading. We then used this to compute expected PI scores. In SI Fig S5d-i, we repeated this analysis by jointly varying the range and baseline shift of curve governing the probability of bump jumps. We used 51 evenly spaced values between 0 and 1 for both the range g and the baseline shift b, subject to the constraint $g \in [0, 1-b]$. Note that in *SI: Reinforcement learning framework*, we use the symbols G_J and B_J to refer to the range and baseline shift of the bump jump curve, rather than g and b (used here for notational simplicity).

Circuit model summary. In Figs 4-5, we constructed and simulated a circuit model that could implement and modify the parameters of a prestructured behavioral policy. This model is described in detail in *SI: Reinforcement learning framework*. Briefly, we model a fly that can fixate and saccade. The duration of its fixations and the directionality of its saccades are determined by three populations of action neurons that receive phase-shifted input about the fly's current heading (from a population of compass neurons) and input about the fly's goal heading (from a population of goal neurons). Both the duration of fixations and the directionality of saccades are modulated by the strength of the goal heading, and by the fly's current bump heading relative to this goal heading. The location and strength of the goal heading are determined by a set of plastic goal weights $\vec{\omega}_G$ that can change over time based on the fly's current bump heading and current level of reinforcement. We additionally account for the stability of the heading itself, which is dictated by a set of plastic compass weights $\vec{\omega}_C$ that change over time based on the fly's current bump heading; these weights capture the net inhibition from ring neurons onto compass neurons and thereby determine the probability that the HD bump will jump between locations that correspond to symmetric views of the visual scene. Bump jumps were assumed to occur immediately following a saccade. Both sets of weights were updated during fixations using simple Hebbian-like plasticity rules (see Equations 37-39 in *SI: Reinforcement learning framework*). We assume that the compass weights, $\vec{\omega}_C$, can be modified continuously, regardless of the presence or absence of reward/punishment. In contrast, we assume that the goal weights, $\vec{\omega}_G$, can only be modified in the presence of reward/punishment. Algorithms 3-4 show how we implemented this model.

To illustrate the behavior of the circuit model, as shown in Figs 4-5, we partitioned heading space into $N = 32$ evenly-spaced headings between $-180°$ and $180°$ (below, we will index these headings as $\theta_i, i \in [1, N]$). To mimic the experimental setup, we simulated two safe zones and two danger zones, each spanning $90°$ of heading space. We defined the centers of the safe zone to be at the arena headings $\theta_A = \{90°, 270°\} = \{\theta_9, \theta_{25}\}$. We used this model to simulate a period of training in which the compass and/or goal weights were evolving over time, and to probe performance and behavior given a fixed set of compass and goal weights. Each training period consisted of a series of iterations, each consisting of a single saccade and a single fixation. During simulations in which the compass and/or goal weights were changing over time, we subsampled periods of fixation into 100ms increments, and iteratively updated the weights for each increment. Weights were not updated during saccades. The duration of fixations and directionality of saccades were determined by the current goal weights, as described above (see Eq. 34 in *SI: Reinforcement learning framework* for more details), and the sizes of saccades were sampled from a lognormal distribution with parameters $\varphi_S = 3.89$ and $\sigma_S = 0.54$ (matched to the values that were fit from data). During probe periods in which the weights remained fixed, there was no simulated reward or punishment. During training periods, the model fly received a reward of $+1$ per unit time when in the safe zone, and a reward of -1 per unit time when in the danger zone.

Prestructured behavioral policy. Fig 4c illustrates the behavioral policy whose heading-dependent structure is guaranteed by the multiplicative operation performed by the three populations of action neurons. We used a profile of goal weights $\omega_G(\theta) = \vec{\omega}_0 \equiv -.2\cos(\theta - \theta_1) - .6\cos^2(\theta - \theta_5) + .5\cos^3(\theta - \theta_2) + .5\cos^4(\theta - \theta_{10})$, normalized to the range $[0, 1]$. As discussed above, we assumed a cosine profile of compass activity, also normalized to the range $[0, 1]$ and centered on $-90°$. The net output of the action neurons was determined by multiplying the profile of goal

weights by the phase-shifted compass activity (phase shifted by $-90°$, $180°$, and $90°$) and summing the output. We repeated this calculation for each possible circular shift of the goal weights to compute the net output as a function of current relative to goal heading. For each possible shift, we computed the probability of CW/CCW saccades and the duration of fixations as outlined in *SI: Reinforcement learning framework*.

Goal weights control location and strength of goal heading. SI Fig S7a-e illustrates the relationship between the profile of goal weights and the gain of fixations and saccades. We compared two weight profiles: $\omega_G(\theta) = .5(1+\cos(\theta-\theta_9))$ ('profile 1') and $\omega_G(\theta) = \vec{\omega}_0$ ('profile 2', given above and normalized to the range $[0,1]$). We scaled the range of each of these profiles by a multiplicative scaling factor g, and we centered the resulting profile at a value of 0.5 (i.e., $\vec{\omega}_{\text{scaled}} = g\vec{\omega} + (1-g)/2$. We used 10 evenly spaced values of g between 0.1 and 1. For each value of the scaling factor, we performed the same calculation described above in *Modeling Methods: Prestructured behavioral policy*, and we measured the gain in the heading-dependent profiles of fixation duration and CW saccade probability to be the difference between the maximum and minimum values of these profiles; these values were shown in SI Fig S7d. We then repeated this analysis, fixing the range of each profile at 0.5, and varying the baseline shift b of each profile. We used 10 evenly-spaced values of b between 0 and 0.4; the results are shown in SI Fig S7e.

Compass weights control the probability and coherence of bump jumps. SI Fig S7f-j illustrates the relationship between the profile of compass weights and the properties of bump jumps. We compared two weight profiles: $\omega_C(\theta) = .5(1+\cos(\theta+\theta_9))$ ('profile 1', chosen such that the most stable heading was aligned with the goal heading used above; see profile 1 for $\omega_G(\theta)$ above), and $\omega_C(\theta) = \vec{\omega}_0$ ('profile 2', normalized to the range $[0,1]$). We again scaled the range of each of these profiles by a multiplicative scaling factor that took 10 evenly spaced values between 0.1 and 1. For each value of the scaling factor, we computed the probability of a jump using Eq. 27; in this way, the probability of a jump at any given orientation θ depends on the difference in compass weights between θ and $\theta + 180°$. In the upper panel of SI Fig S7i, we reported the probability of a jump when averaged over all headings. In the lower panel of SI Fig S7i, we measured coherence by summing the relative probability that the bump would jump toward, rather than away from, the most stable bump heading. We computed this by summing $p_{\text{jump}}(\theta+180°) - p_{\text{jump}}(\theta)$ over the range of θ values within $\pm 90°$ of the most stable bump heading. We then repeated this analysis, fixing the range of each profile at 0.5, and varying the baseline shift s of each profile. We again used 10 evenly-spaced values of s between 0 and 0.4; the results are shown in SI Fig S7j.

Weight updates. In Figs 4-5, we simulated the behavior of the circuit model while training the compass weights alone (Fig 4g), and while jointly training the heading and goal weights (Fig 5a-f). The Hebbian-like learning update that we used is given in Eq. 39, and example updates to weight profiles are illustrated in Figs 4g and 5a. These illustrations were generated using the base profiles $\omega_C(\theta) = \omega_G(\theta) = 0.4 + .125(-.2\cos(\theta-\theta_1) - .4\cos^2(\theta-\theta_5) + 1)$, with $\alpha_C = \alpha_G = 0.05$ and $\theta_C = \theta_8$.

Compass weights tether to goal weights during learning. Fig 4g illustrates how the compass weights evolve over time for a fixed profile of goal weights. We initialized the compass weights to $\omega_C(\theta) = .3(.2\cos(\theta-\theta_6) + .4\cos(\theta-\theta_{12})^2 + 1)$, and we used a fixed set of goal weights, given by $\omega_G(\theta) = .5(1+\cos(\theta-\theta_9))$. We simulated the model for 400 simulated seconds of training. The third panel of Fig 4g shows the histogram of bump headings accumulated over the duration of training. The bottom panel of Fig 4g shows the evolution of the weight profile during the first quarter of training.

Circuit model with fixed weights reproduces structure of fly behavior. In Fig 4h, we illustrated the output of the circuit model using fixed sets of compass and goal weights. We used the weight profiles $\omega_C(\theta) = .5(1+\cos(\theta-\theta_9))$ and $\omega_G(\theta) = .5(1+\cos(\theta+\theta_9))$ to simulate the output of the circuit model in the absence of any training (i.e., assuming these weights remain fixed for the duration of the simulation); these profiles are shown in the upper row Fig 4h. We then simulated the behavior of 1000 model flies over 500 (where, as described above in *Modeling Methods: Circuit model summary*, each iteration consisted of a single fixation and single saccade). To compute the simulated $\Delta F/F$ profile as a function of arena and bump heading (shown in the lower right panel of Fig 4h), we accumulated all iterations for which a given model fly visited each of the N arena headings. We then averaged the bump profile (carried in the population of compass neurons) across all visits to the given arena heading, weighted by the duration of fixation at that heading. This produced a matrix of $\Delta F/F$ values specified for combination of each arena and bump heading. We computed this matrix for each of the 1000 model flies, aligned each matrix to the goal heading and corresponding arena heading of each fly, and then we averaged this matrix across flies. To

compute the average probability of a bump jump, average duration of fixations, and average direction of saccades conditioned on either bump heading or arena heading (shown in the middle panels of Fig 4h), we accumulated all iterations for which a given model fly visited each of the N arena headings during the first 100 iterations, and we averaged the fraction of iterations that the bump jumped, the duration of fixations, and the directionality of saccades taken at that particular heading. We again repeated this for each of the 1000 model flies, aligned to the preferred heading of each fly, and averaged the results across flies.

Training the compass and goal weights of the circuit model. Fig 5 illustrated the behavior of the circuit model when the compass and goal weights are changing over time, shown for a single simulated fly (panels 5b,d,e) and summarized across 1000 model flies (panels 5c,e,f). To illustrate the behavior of the single fly, we initialized the weights to be $\omega_C(\theta) = .72 - .3(.2\cos(\theta - \theta_6) + .4\cos(\theta - \theta_{12})^2 + 1)$ (as above), and $\omega_G(\theta) = .3(.2\cos(\theta - \theta_6) + .4\cos(\theta - \theta_{12})^2 + 1)$. Given that these initial weights are very weak and unstructured, we simulated learning for a longer period of 2000 simulated seconds. The left column of Fig 5b shows the temporal evolution of $\vec{\omega}_C$ and $\vec{\omega}_G$ during the training period; the right column of the same panel shows the corresponding evolution of the motor drive. To compute the arena and bump residencies shown in Fig 5d, we froze the weights before training, after 500 simulated seconds, and at the end of training. We used these frozen weights to simulate 1000 seconds of behavior, and we used this simulated behavior to compute the fraction of time that the fly and the bump spent at different headings. Fig 5e shows the temporal evolution of the range of the goal weights (measured as the difference between maximum and minimum weights values), the minimum angular distance between the goal heading and the center of the safe zone, and the PI score over the course of training; these values were calculated over consecutive windows of 100 simulated seconds with an overlap of 50 simulated seconds.

To illustrate the impact of training across many model flies, we simulated learning for 200 seconds, closely matched to duration of pairs of training trials used in the learning assay. We used weight profiles of the same fixed form $g/2(1 + \cos(\theta - \theta_i)) + (1 - g)/2$, but we randomly varied the location θ_i and the gain g for both the compass and the goal weights. The left column of Fig 5c summarizes the goal heading and the most stable bump heading for each model fly before training (upper row), one quarter of the way into training (middle row), and three quarter of the way into training (lower row). The right column of Fig 5c summarizes the range of compass and goal weights at these same time points. For each of these three time points, we froze the compass and goal weights, and we used these frozen weights to simulate 1000 seconds of behavior and calculate PI scores. The change in PI score was used as the color code in Fig 5c. As described in the previous paragraph, we computed the temporal evolution of the range of the goal weights, the minimum angular distance between the goal heading and the center of the safe zone, and the PI score during the training period for each individual model fly, and we displayed the average and standard error across flies in the right column of Fig 5e. These measures were computed using consecutive windows of 20 simulated seconds with an overlap of 10 simulated seconds. We then separated flies into two groups according to whether the range of their goal weights was greater or less that 0.5 (high and low range, respectively, in the upper panel of Fig 5f), and separately according to whether their initial goal heading was within the safe or danger zone (preference for safety or danger, respectively, in the lower panel of Fig 5f). See *Analysis Methods: Measuring the distance to safety* and *Analysis Methods: Measuring the degree of structure in the behavior* for analogous measures computed for behavioral data. SI Fig S8a-c shows these same analyses, separated by both initial range and initial preference. SI Fig S8d shows the cumulative distribution of incoherent weight updates for these same groups. To measure this, we first determined which safe zone (left versus right) minimized the angular distance to the final internal goal heading, and we used the center of that safe zone as the reference goal. Then, for each training iteration (where, as before, a single iteration includes a single saccade and single fixation), we computed the net weight update, and we determined whether (and by how much) the net update moved the internal goal heading closer or further from the reference goal relative to the previous iteration. We measured the fraction of incoherent weight updates as the fraction f of iterations for which the difference in net update between successive iterations changed sign (indicating that the goal heading was shifting closer then further from the reference goal, or vice versa) and exceeded a magnitude of 0.1. We reported the fraction of coherent updates, $1 - f$, in SI Fig S8d.

SUPPLEMENTAL INFORMATION

Linking the Conceptual Model to Known Anatomy

Our model (SI Fig S6a) builds on the many conceptual ideas and models that have been proposed for the CX in recent years. Several of our assumptions are based on physiological studies of different CX neuron types during visual stimulation and behavior. Importantly, although not all the key features of our circuit model have physiological support, they are all inspired by the known anatomy and connectivity of the CX. Rather than incorporating all the known details of CX connectivity, however, we greatly simplified and abstracted the circuit in order to focus on the key computations that we believe underlie the fly's behavior in the visual learning paradigm. We now discuss the many simplifications that we made, and summarize what is known of several neuron types and network motifs that are likely to play a role in relevant circuit computations.

The flexible mapping from visual scene to HD representation: Recurrent connections between ring neurons and EPG neurons. The HD system is tethered to its sensory surroundings by multiple ring neuron classes —most with tens of ring neurons each— that carry information about sensory cues, such as visual features [65–67, 101], polarized light patterns [102–104], and wind direction [105]. Most of these neurons are thought to be GABA-ergic and inhibitory [106]. Some visual-feature-sensitive ring neurons have spatiotemporal receptive fields [67], and are thus likely to be sensitive to how a visual feature moves across the fly's eyes, a factor that we ignored in our model. Sensory ring neurons are connected all-to-all to other ring neurons of the same class, and, in some cases, across classes as well, and most sensory ring neurons receive feedback from the compass (EPG) neurons [34]. These motifs may ensure that the fly's HD representation tethers to the strongest cues available [34], which we assume to be less relevant in a visual setting with four identical horizontal bars. Plasticity in the synapses between ring and compass neurons has been hypothesized to create a flexible mapping between sensory cues and the HD representation [60, 64], an idea similar to one proposed for the rodent HD system [107] (SI Fig S6b). Recent experimental results strongly support this idea of plasticity between visual inputs and compass neurons [29, 30]. As part of a model proposed in one of these studies [29], we assumed that this plasticity depends on an inhibitory Hebbian-like rule that relies on correlated activity between visual and compass neurons, and results in changes in the depth of inhibition that compass neurons receive at different angular orientations in their surroundings [30]. Plasticity in the EB may involve nitric oxide signaling [108] and motor-state-dependent neuromodulation, which we did not explicitly model in this study. There are multiple sources of neuromodulation in the EB, including the ExR2 dopaminergic neurons (SI Fig S6b) , which receive inputs in the LAL and project to the BU and EB (see Figure 14 and associated figure supplements in [34]) and have been linked to circadian changes in locomotion [109].

Rather than explicitly modeling plasticity in the mapping between ring and compass neurons, our conceptual model captures the impact of this plasticity on HD bump dynamics at different orientations. Specifically, several experimental studies [19, 29, 30] have reported instabilities in the EPG bump's offset relative to the fly's surroundings in symmetric visual settings. Two visually indistinguishable headings are likely to evoke similar ring neuron population activity, making both EB locations corresponding to those headings viable for the bump to occupy. Which one of those locations the bump resides in would depend on the relative strength of the ring-neuron-to-EPG mapping in those two EB locations. In our behavioral paradigm, we found that some EB locations were more likely to feature bump jumps (see Fig 3k-l), and that these bump jumps tended to be 180° in magnitude (see Fig 3i), reflecting the symmetry of the scene (note that the vertical span of some ring neurons' receptive fields may be large enough to evoke responses to horizontal bars at both high and low elevations, making the scene weakly symmetric at 90° from the perspective of those inputs to the compass neurons, and triggering a few 90° bump jumps as well (small peak in Fig 3i, bottom panel)). In addition, we noticed that the probability of bump jumps, which we expect to reflect the strength or weakness of synapses from inhibitory ring neurons onto the local EPG population, depended on the relative residency of the EPG bump in those locations for the same visual setting. This motivated our model's assumption that the presence of the HD bump in an EB location triggers plasticity between visual and compass neurons. This allowed connection strengths to be modulated during fixations, rather than just during saccades, in contrast to [29], where plasticity is dependent on the fly's angular velocity. Thus, the more time that the EPG bump spends in a specific EB location for a given heading relative to the visual scene, the more strongly the ring neurons tether it to that location in the future. We captured this, not by explicitly modeling ring neuron and EPG interactions, but by modifying the strength or 'gain' of the HD representation at different EB locations. Additionally, rather than directly modeling the impact of inhibitory interactions from ring neurons onto EPG neurons, and the consequence of these experience-dependent interactions on bump dynamics, we simply assumed that the relative, experience-dependent strength of the HD representation in 180°-opposite EB positions determines the probability of an EPG bump jump between them. Note that these differences in the strength of summed ring neuron inhibition

onto EPG neurons at $180°$-opposite EB locations would not necessarily lead to EPG bump amplitude differences at those locations, because of other sources of broad feedback inhibition onto EPG neurons in both the EB and the PB [34, 43, 63, 110].

Maintaining and updating the HD representation: A ring attractor circuit involving EB and PB neurons. There are $\sim$48 EPG neurons, which each occupy one of 16-18 compartments in the EB and PB [34, 111]. In our model, we assumed that the sinusoidal HD representation is carried by 32 EPG-like neurons, each with distinct HD tuning equally spaced across $360°$. Although EPG neurons have diffuse EB arbors and contact other EPGs within and even across EB compartments (wedges), which may enable a relatively continuous HD representation in the EB, it is more likely that the $360°$ of the fly's HD are represented discretely in the 16-18 compartments of the EB and PB (SI Fig S6c).

The dynamics of the HD representation match those produced by ring attractor networks [43, 58, 61]. These dynamics depend not just on sensory inputs, but also on self-motion input from the PEN_a (SI Fig S6c) and other columnar neurons that link the EB and PB [43, 58, 62]. These self-motion inputs are also important for the mapping of visual scenes onto the EPG population [29]. In our model, the HD representation was entirely driven by visual input, and we ignored recurrent connections involving the PEN and PEG neurons, as well as the intra-EB connections between different EPG neurons. These connections could, in effect, tether the EPG bump more strongly to locations near its current location, reducing the probability of bump jumps even when favorable based purely on the relative strength of connections from ring neurons.

Neurons that reformat the HD bump and link the PB to the FB Our model assumed that the EPG population's HD representation is always sinusoidal, but we did not explicitly include connections from EPG neurons onto the broadly arborizing PB neurons that are thought to be key to maintaining this shape, the $\Delta 7$ neurons [34, 63] (SI Fig S6c). Both the EPG neurons and $\Delta 7$ neurons contact a wide range of columnar neurons in the PB. Some of these neurons project back to the EB, but most are FB columnar neurons [34], such as the PFNs and PFRs [106, 112, 113] (SI Fig S6d). Many of these FB columnar neurons receive additional inputs in other CX structures [34, 63, 114–116]. In combination with neuron-type-specific anatomical phase shifts in their projection patterns from individual glomeruli in the PB to columns of the FB [34, 63, 115] (see SI Fig S6d for illustrative examples), these inputs likely allow the PB-FB columnar neuron types to participate in vector computations that transform the HD representation in different ways [34, 63, 115]. This may be highly relevant to the computations that flies use when navigating over long distances in natural conditions [45, 49]. In such natural settings, flies would likely need to select actions to maintain a particular traveling direction ('goal heading') rather than to maintain a specific head direction ('goal HD'). Recent conceptual insights from the connectome [34] and, in parallel, confirmatory evidence from physiological experiments [63, 115] suggest that some FB columnar neurons use translational self-motion cues to transform HD into an explicit representation of traveling direction (heading). At the end of this Supplemental section, we discuss how a potential circuit mechanism to learn and express a heading (traveling direction) preference that is robust to perturbations from wind might be implemented in FB circuitry. However, in a head-fixed preparation in which the fly only controls and receives visual feedback for its angular movements, the distinction between HD and heading is less relevant. Thus, although we use the term 'heading' in the main text and below, our model did not incorporate mechanisms that would be necessary for this flexible behavior to operate in the space of heading rather than HD.

Learning and storing a goal heading: Candidate neurons and circuitry in the FB. Recent studies have proposed simple conceptual models for how goal headings might be stored in the strength of synaptic connections between neurons of the FB [32, 34]. An entirely different model for visually-guided homing is that heading-dependent views or visual snapshots are stored in the strength of synapses between visually-responsive Kenyon cells and mushroom body output neurons (MBONs) in the MB [71, 72, 74, 117]. There is, as yet, little experimental evidence for these ideas (but see [73, 118]). However, there is evidence that flexibility in goal headings depends on visual input to ring neurons [43] and on output from EPG neurons [28, 31]. Further, this flexibility does not involve changes in the mapping between the visual scene and the EPG HD representation [31], suggesting that goal headings are stored downstream of the EPG neurons. The behavioral genetics evidence implicating FB neurons in a visual learning task that inspired ours [27] suggest that goal headings are stored in plastic synapses between neurons in the FB.

We assumed that the goal heading is stored in the strength of synapses from hypothesized tangential motor state neurons to putative columnar 'goal neurons' (SI Fig S6e-f). Such a mechanism would be metabolically efficient, allowing goal neurons to be activated into a goal-heading pattern specifically when the fly is moving, but not

otherwise. We further assumed that these synaptic strengths are modified through the action of neuromodulatory tangential FB neurons, which, we assume, deliver reinforcement signals shaped like the current heading bump (SI Fig S6e). There is already some physiological evidence for reinforcement signals in the FB [68], although the neuronal players involved are as yet unknown. A prime candidate for such a role is the ventral FB dopaminergic neuron (DAN) type. MB DANs are known to carry reinforcement (and movement) signals and be involved in associative learning in that brain region [119–123], and a subset of FB DANs —potentially the ventral FB DANs, such as FB2A, FB4L or FB4M, rather than the dorsal FB DANs that have been associated with sleep- and nutrient-related signaling— may well perform similar functions. Interestingly, most tangential neurons that innervate the ventral FB receive local input from columnar FB neurons —including subtypes of PFN, PFR, vΔ and hΔ neurons— near their presynaptic sites in the FB [34], suggesting that any reinforcement signals they might carry would likely be locally shaped by heading input, as required by our model. The ventral FB DANs as well as other tangential FB neurons send their outputs to other FB tangential neurons and to many columnar neurons, including the hΔ, vΔ and FC neurons [34]. These connectivity patterns match our model's proposal that the reinforcement signal acts on synapses between motor-state-carrying tangential neurons and columnar goal neurons (SI Fig S6f). Indeed, some of the ventral FB DAN's tangential targets receive inputs in the motor-activity-related LAL and NO, making these target neurons ideally situated to carry motor state information [34]. At least one type of FB tangential neurons —likely a subtype of FB2B [34]— has been implicated in visual learning [27], receives DAN input from FB2A neurons [34], and also has strong motor-state-dependent activity [124].

Implementing a policy of a fixed form: Phase shifts of CX output neurons. A remarkable feature of several CX columnar neurons is the precision of their projection patterns within different CX structures. In particular, most CX columnar neurons that project from the PB to either the EB or the FB show precise 'phase shifts' between their localized arbors in the different structures. These phase shifts are computed relative to the projected position of the EPG bump in different structures. For example, PEN neurons whose arbors overlap with EPG neurons in the PB project to a location in the EB that is shifted by 45° relative to their input EPGs [113]. In the case of the PEN_a neurons, this phase shift has been proposed to allow self-motion-derived angular velocity input to shift the position of EPG population activity in the EB [58, 62]. Our model relies on the projection patterns of the PFL2 and PFL3 neurons, which show phase shifts of 180° and 90° respectively between their PB and FB arbors (SI Fig S6d,f; note also that PFL2 neurons project to both left and right LALs). Our proposal for how these phase shifts might enable action selection is based on ideas proposed in [34] and, in the case of the PFL3 neurons, also shares similarities with a model that was proposed for path integration in the sweat bee [59]. As shown in Fig 4c, the phase shifts automatically enable a multiplication of a phase-shifted version of the fly's current heading with its goal heading. For the PFL2 neurons, the 180° shift means that the product of this multiplication peaks when the fly is heading in exactly the opposite direction to the goal heading. If activity in the PFL2 neurons modulates drift rate within neurons that control fixation, as we propose, then the result of peak activity would be high drift rate that results in shorter fixation and transitions to turning. Thus, PFL phase shifts provide a potential mechanism to directly induce actions that would steer the bump towards a goal heading, allowing learning to work in the lower-dimensional space of merely updating the goal weight vector, rather than the space of actions necessary to direct the fly to its goal. In addition to their direct projections onto descending neurons (DNs) [32, 34] that send motor commands to the thorax [125–127] (SI Fig S6f), the PFL2 and PFL3 neurons also converge onto LAL neurons that themselves project onto DNs [34]. These projection patterns justify our modeling assumptions regarding how heading-tethered PFL (action neuron) activity is converted into directional motor commands.

A potential circuit mechanism to learn and travel in the direction of a true goal heading. As we discussed above, our model ignores the distinction between HD and heading for the purposes of modeling the fly's behavior in our paradigm. However, we believe that this paradigm exploits a mechanism that the fly uses during dispersal and long-range navigation in more natural settings [45, 49]. The FB's circuits could provide the requisite mechanism for the fly to travel in a specific direction, that is, to learn and then progress towards a goal heading. In essence, to produce appropriate movements, PFL neurons would need to receive FB input that has already accounted for the effects of different head-body angles and perturbations such as wind input through appropriate vector computations [34, 63].

For example, imagine a fly attempting to travel northeast on a windy day. At every moment in time, the fly's total translational velocity ('TV') vector will be the sum of two components: the influence of the wind and the fly's self-generated movement, which we assume is in the same direction as its head direction (i.e., the fly's head direction matches its body direction). That is, $\mathrm{TV}(t) = \mathrm{Wind}(t) + \mathrm{Fly}(t)$. In this case, if the wind were blowing the fly east, the fly would have to fly north with the same velocity as $\mathrm{Wind}(t)$ to maintain a northeast heading.

Here we assume that $\mathrm{Fly}(t)$ is the HD bump scaled by the fly's forward flight velocity, and that $\mathrm{TV}(t)$ is the fly's total translational velocity vector, carried by hΔB neurons [63, 115]. With these two signals, the fly could compute the allocentic wind direction (i.e., $\mathrm{Wind}(t) = \mathrm{TV}(t) - \mathrm{Fly}(t)$). Next, to compute the desired, or 'goal' (G), head direction, the fly would have to subtract the instantaneous wind direction from the goal translation vector (i.e., $\mathrm{HD}_G(t) = \mathrm{TV}_G - \mathrm{Wind}(t)$). Here we assume that TV_G is a vector that is stored in the FB that encodes the fly's desired travel direction. Once the desired head direction ($\mathrm{HD}_G(t)$) has been computed, the PFL2/3 neurons could generate goal-directed motor commands as described above. The complexity of this computation arises because the PFL2/3 neurons are thought to inherit the fly's HD in the PB, which would prevent them from directly comparing the fly's instantaneous TV ($\mathrm{TV}(t)$) to its goal TV (TV_G). Instead, to accurately account for the influence of the wind, the PFL2/3 neurons would have to receive the desired HD vector ($\mathrm{HD}_G(t)$) as input in the FB. In this way, the PFL2/3 neurons could compare the fly's current HD to its goal HD. This conceptual model demonstrates that the fly CX could, in principle, account for external perturbations when selecting appropriate actions. Alternatively, instead of storing a desired travel direction, the fly could store a desired HD. Doing so could allow the fly to set an approximately accurate course under some circumstances. For example, if the wind consistently blew the fly east, maintaining a north head direction would yield an overall northeast travel, on average, but this mechanisms would not allow the fly to account for external perturbations like the wind. Future physiological recordings during behavior are required to assess which of these two classes of goal vectors the fly may be using and in which contexts.

Reinforcement learning framework

In this study, we used a reinforcement learning framework to explore how the fly's behavioral modes, namely fixations and saccades, should be structured as a function of the fly's heading and updated based on the fly's experience. We first segmented behavior into these two modes, and used the variability in these modes to inform the structure and adaptable control parameters of a behavioral policy, as described in *Analysis Methods: Inferring the structure of a behavioral policy*. Here, we build an agent that uses this policy to structure its behavior as a function of its orientation within its visual surroundings. We then used reinforcement learning to train the control parameters of this policy according to different objectives.

In the main text, we considered two primary objectives: (i) maintaining a preferred orientation with respect to a visual pattern in the arena, and (ii) maintaining a preferred heading in the arena. We used these two objectives to explore how behavior might have been structured over evolutionary timescales, and thus compared the results of this learning to naive fly behavior. As shown in Fig 2f-g, these two objectives give rise to control parameters that are sinusoidally structured as a function of the fly's arena heading. We showed that control parameters obtained by the second of these objectives, when mediated by an unstable internal representation of heading and controlled with respect to an internal goal heading, produces behavior that qualitatively resembles fly behavior. With such a policy that is structured with respect to this internal goal heading, the fly need only shift the location of the goal heading to adapt to new surroundings. We thus refer to this policy as a "prestructured internal policy" that specifies a set of actions that are tethered to the fly's internal representation of heading relative to an internal goal heading. Below, we show how reinforcement learning can be used on shorter timescales to shift the location of the goal heading based on experience. Finally, we illustrate how a simplified circuit model can be used to implement this prestructured policy, and how plasticity in two sets of weights within the circuit can be used to update the goal heading and most stable bump heading over time.

In what follows, we first outline the general form of the policy and the training algorithm. We then show how the control parameters of this policy can be learned using a policy gradient algorithm with function approximation; we use this approach to study how the control parameters should be structured as a function of heading to maintain a preference for a specific visual pattern or a specific heading. We then assume that the structure in these control parameters is built-in to the policy and maintained with respect to a single goal heading, and we show how a policy gradient algorithm can be used to update the location of this goal heading while otherwise maintaining the structure of the policy. Finally, we outline the circuit implementation of this prestructured policy, and we show how Hebbian-like plasticity can be used to implement the learning process.

General architecture

Policy. We consider a general scenario in which the behavior of an agent (model fly) is governed by a stochastic policy $\pi(\dot{\theta}, \Delta t|\theta)$. This policy determines the probability of maintaining an average angular velocity $\dot{\theta}$ over a duration of time Δt given an initial heading θ, and thereby determines the probability of generating a change in heading

$\Delta\theta = \dot{\theta}\Delta t$. We will use Δt to denote the duration of a single action that is sampled from the policy; depending on the scenario, this can correspond to the entire duration of a saccade, the entire duration of a fixation, or the duration of a sampling event within a fixation. The policy is parameterized by a set of fixed parameters $\vec{\beta}$ that do not change over time, and a set of flexible parameters $\vec{\omega}$ that can be modified through experience. For notational simplicity, we will introduce these parameters as they become necessary. When writing a conditional distribution $P(A|B)$, we will explicitly denote the dependence on $\vec{\omega}$ when it exists (i.e., $P(A|B;\vec{\omega})$), and will assume implicit dependence on $\vec{\beta}$ when it exists.

We decompose behavior into two different behavioral modes: fixations ('F'), and saccades ('S'), such that the policy can be written as:

$$\begin{aligned}\pi(\dot{\theta}, \Delta t|\theta) &= P(\dot{\theta}, \Delta t|\theta, F)P(F|\theta) + P(\dot{\theta}, \Delta t|\theta, S)P(S|\theta) \\ &= P(\Delta t|\dot{\theta}, \theta, F)P(\dot{\theta}|\theta, F)P(F|\theta) + P(\Delta t|\dot{\theta}, \theta, S)P(\dot{\theta}|\theta, S)(1 - P(F|\theta))\end{aligned} \tag{1}$$

Here, we have used the constraint that $P(F|\theta) + P(S|\theta) = 1$, and we have incorporated the observation that the duration of each mode can be conditioned on the angular velocity (see *Analysis Methods: Characterizing fixation properties* and *Analysis Methods: Characterizing saccade properties*).

Fixation policy. We assume that fixations are generated in a timepoint-by-timepoint manner via a drift diffusion process with integrated signal ξ; when this signal crosses a fixed threshold, the fixation is terminated and a saccade is initiated. The probability of maintaining a fixation thus depends on ξ:

$$\pi(\dot{\theta}, \Delta t|\theta, \xi) = P(\Delta t|\dot{\theta}, \theta, F)P(\dot{\theta}|\theta, F)P(F|\theta, \xi) + P(\Delta t|\dot{\theta}, \theta, S)P(\dot{\theta}|\theta, S)(1 - P(F|\theta, \xi)) \tag{2}$$

We assume that the integrated signal ξ is updated in time increments of δt, and we assume that there is no change in heading during this time increment. This allows us to define:

$$P(\Delta t|\dot{\theta}, \theta, F) = \delta(\Delta t - \delta t) \tag{3}$$

$$P(\dot{\theta}|\theta, F) = \delta(\dot{\theta}) \tag{4}$$

During fixations, the integrated signal ξ is updated by an amount $\Delta\xi$ that is determined by the drift diffusion process. We model this process with a fixed spread η_F^2 and a heading-dependent drift rate $\nu_F(\theta;\vec{\omega})$ that is parameterized by the flexible parameters $\vec{\omega}$:

$$P(\Delta\xi|\theta, \xi; \vec{\omega}) = \mathcal{N}(\nu_F(\theta;\vec{\omega})\delta t, \eta_F^2\delta t) \tag{5}$$

This update will terminate the fixation and result in a saccade if the net signal $\xi + \Delta\xi$ crosses a fixed threshold a_F. Note that this produces an inverse Gaussian distribution of fixation durations ΔT, with average durations $a_F/\nu_F(\theta;\vec{\omega})$ (where here, we use ΔT to denote the duration of the entire fixational event, computed in increments of δt):

$$P\left(\Delta T|\dot{\theta}, \theta, F; \vec{\omega}\right) = \mathrm{IG}\left(\Delta T; a_F/\nu_F(\theta;\vec{\omega}), a_F^2/\eta_F^2\right) \tag{6}$$

This allows us to write the probability of fixating as:

$$\begin{aligned}P(F|\theta, \xi; \vec{\omega}) &= \int P(F|\Delta\xi, \theta, \xi)P(\Delta\xi|\theta, \xi)d\Delta\xi \\ &= \int^{a_F} P(\Delta\xi|\theta, \xi)d\Delta\xi \\ &= \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{a_F - (\xi + \nu_F(\theta;\vec{\omega})\delta t)}{\sqrt{2\eta_F^2\delta t}}\right)\right)\end{aligned} \tag{7}$$

In the presence of heat, one can include second, reflexive drift process that can short-circuit the termination of a fixation (see *Analysis Methods: Characterizing fixation properties* for motivation). To illustrate how this could be carried out, we consider a process that is governed by an integrated signal ξ_R that obeys the same dynamics as above, with the same spread η_F^2 but with a fixed drift rate ν_R. A fixation is terminated whenever either ξ or ξ_R cross the fixed threshold a_F. The probability of fixating is thus determined by:

$$P(F|\theta,\xi,\xi_R;\vec{\gamma}) = \min\Big\{P(F|\theta,\xi;\vec{\gamma}), P(F|\theta,\xi_R)\Big\} \tag{8}$$

where

$$P(F|\theta,\xi_R) = \frac{1}{2}\left(1+\text{erf}\left(\frac{a_F-(\xi_R+\nu_R\delta t)\Theta(\text{heat})}{\sqrt{2\eta_F^2\delta t}}\right)\right) \tag{9}$$

Here, $\Theta(\text{heat})$ is a heaviside function that takes a value of 1 if there is a perceived heat, and 0 otherwise. Note that in the absence of perceived heat, $P(F|\theta,\xi;\vec{\gamma})$ will always be less than $P(F|\theta,\xi_R)$, and will thus determine the probability of fixating through Eq (8).

Saccade policy. We assume that saccades are initiated in a ballistic manner following the termination of a fixation, such that taking a saccade results in an abrupt change in heading $\Delta\theta = \dot{\theta}\Delta t$ over a duration of time Δt.

We assume that the directionality of saccades is controlled via a heading-dependent directional bias $d_S(\theta;\vec{\omega})$ that is parameterized by the flexible parameters $\vec{\omega}$; $d_S(\theta;\vec{\omega})$ specifies the probability of initiating a rightward, or CW, saccade at heading θ. We then assume that the angular speed of a saccade $|\dot{\theta}|$ is drawn from a lognormal distribution with parameters φ_S, σ_S^2. Together, this results in the following distribution over angular velocities $\dot{\theta}$:

$$P(\dot{\theta}|\theta,S;\vec{\omega}) = \left[d_S(\theta;\vec{\omega})\,\text{sgn}(\dot{\theta}) + \left(\frac{1-\text{sgn}(\dot{\theta})}{2}\right)\right]\text{logn}(|\dot{\theta}|;\varphi_S,\sigma_S) \tag{10}$$

We further assume that the duration of saccades, analogously to the duration of fixations, can be generated via a drift diffusion process. Here, we assume that the drift rate is not flexible, but depends on the angular speed of the saccade:

$$P(\Delta t|\dot{\theta},\theta,S) = \text{IG}(\Delta t; a_S/\nu_S(|\dot{\theta}|), a_S^2/\eta_S^2) \tag{11}$$

where $\nu_S(|\dot{\theta}|)$ is well-captured by a sigmoidal function of $|\dot{\theta}|$ (see *Analysis Methods: Characterizing saccade properties*). We can now specify the full policy and its parameter dependence:

$$\pi(\dot{\theta},\Delta t|\theta,\xi;\vec{\omega}) = P(\Delta t|\dot{\theta},\theta,F)P(\dot{\theta}|\theta,F)P(F|\theta,\xi;\vec{\omega}) + P(\Delta t|\dot{\theta},\theta,S)P(\dot{\theta}|\theta,S;\vec{\omega})(1-P(F|\theta,\xi;\vec{\omega}))$$

$$\begin{aligned}
P(\Delta t|\dot{\theta},\theta,F) &= \delta(\Delta t-\delta t)\\
P(\dot{\theta}|\theta,F) &= \delta(\dot{\theta})\\
P(F|\theta,\xi;\vec{\omega}) &= \frac{1}{2}\left(1+\text{erf}\left(\frac{a_F-(\xi+\nu_F(\theta;\vec{\omega})\delta t)}{\sqrt{2\eta_F^2\delta t}}\right)\right)\\
P(\Delta t|\dot{\theta},\theta,S) &= \text{IG}(\Delta t; a_S/\nu_S(|\dot{\theta}|), a_S^2/\eta_S^2)\\
P(\dot{\theta}|\theta,S;\vec{\omega}) &= \left[d_S(\theta;\vec{\omega})\,\text{sgn}(\dot{\theta}) + \left(\frac{1-\text{sgn}(\dot{\theta})}{2}\right)\right]\text{logn}(|\dot{\theta}|;\varphi_S,\sigma_S)
\end{aligned} \tag{12}$$

where $\vec{\omega}$ controls the heading dependence in both the duration of fixations (through the drift rate ν_F) and the directional bias of saccades (through d_S). As noted above, the policy depends implicitly on a set of fixed parameters $\vec{\beta} = [\delta t, a_F, \eta_F, \varphi_S, \sigma_S, a_S, \eta_S]$ that controls inflexible aspects of behavior.

Training. We use an online policy-gradient method to iteratively update the flexible policy parameters $\vec{\omega}$ based on the agent's actions and on the outcome of these actions. In this way, the agent (i) samples an action $[\dot{\theta},\Delta t]$ from its policy based on its current heading θ, (ii) observes the outcome of this action (a change in heading $\Delta\theta = \dot{\theta}\Delta t$, and a heading-dependent sensory response $R(\theta+\dot{\theta}\Delta t)$), and (iii) updates the policy weights $\vec{\omega}$ to modify the probability of taking the same action from the same heading in the future (depending on whether that action led to a good or bad outcome). We use a policy gradient algorithm [16] to update these weights:

$$\begin{aligned}
\Delta\vec{\omega} &= R(\theta+\dot{\theta}\Delta t)\nabla_{\vec{\omega}}\log\pi(\dot{\theta},\Delta t|\theta;\vec{\omega})\Big|_{\theta^*,\dot{\theta}^*,\Delta t^*}\\
&= R(\theta+\dot{\theta}\Delta t)\frac{\nabla_{\vec{\omega}}\pi(\dot{\theta},\Delta t|\theta;\vec{\omega})}{\pi(\dot{\theta},\Delta t|\theta;\vec{\omega})}\Big|_{\theta^*,\dot{\theta}^*,\Delta t^*}
\end{aligned} \tag{13}$$

where θ^*, $\dot{\theta}^*$, and Δt^* denote specific values of the heading, angular velocity, and duration of an action, respectively. We assume that $R(\theta+\dot{\theta}\Delta t)$ is computed directly from changes in sensory experience, rather than from a comparison between sensory experience and expected value. Note that the sensory response effectively acts as the step size, or learning rate, in the update equation for $\vec{\omega}$. As a result, a strong sensory response can result in quick but coarse updates, whereas as weak sensory response will result in slower but finer updates. The policy gradients can then be computed as follows:

$$\begin{aligned}\nabla_{\vec{\omega}}\pi =& P(\Delta t|\dot{\theta},\theta,F)P(\dot{\theta}|\theta,F)\nabla_{\vec{\omega}}P(F|\theta,\xi;\vec{\omega}) + \\ & P(\Delta t|\dot{\theta},\theta,S)\Big[(1-P(F|\theta,\xi;\vec{\omega}))\nabla_{\vec{\omega}}P(\dot{\theta}|\theta,S;\vec{\omega}) - P(\dot{\theta}|\theta,S;\vec{\omega})\nabla_{\vec{\omega}}P(F|\theta,\xi;\vec{\omega})\Big] \\ =& \Big[P(\Delta t|\dot{\theta},\theta,F)P(\dot{\theta}|\theta,F) - P(\Delta t|\dot{\theta},\theta,S)P(\dot{\theta}|\theta,S;\vec{\omega})\Big]\nabla_{\vec{\omega}}P(F|\theta,\xi;\vec{\omega}) + \\ & \Big[P(\Delta t|\dot{\theta},\theta,S)(1-P(F|\theta,\xi;\vec{\omega}))\Big]\nabla_{\vec{\omega}}P(\dot{\theta}|\theta,S;\vec{\omega})\end{aligned} \tag{14}$$

We can use Eq. (12) to further simplify the gradients:

$$\nabla_{\vec{\omega}}P(F|\theta,\xi;\vec{\omega}) = -\delta t\mathcal{N}(a_F;\xi+\nu_F(\theta;\vec{\omega})\delta t,\eta_F^2\delta t)\nabla_{\vec{\omega}}\nu_F(\theta;\vec{\omega}) \tag{15}$$

$$\nabla_{\vec{\omega}}P(\dot{\theta}|\theta,S;\vec{\omega}) = \mathrm{sgn}(\dot{\theta})\,\mathrm{logn}(|\dot{\theta}|;\varphi_S,\sigma_S)\nabla_{\vec{\omega}}d_S(\theta;\vec{\omega}) \tag{16}$$

Further evaluating the policy gradients depends on the form of $\nabla_{\vec{\omega}}\nu_F(\theta;\vec{\omega})$ and $\nabla_{\vec{\omega}}d_S(\theta;\vec{\omega})$, which are determined by the specific implementations that we consider below.

Implementation. In the main text, we considered different variants of this basic framework. The first variant, shown in Fig 2f-g, considers a "flexible" policy in which $\nu_F(\theta;\vec{\omega})$ and $d_S(\theta;\vec{\omega})$ can be modified in a heading-dependent manner via function approximation with a set of weights $\vec{\omega} = [\vec{\omega}_F,\vec{\omega}_S]$. Learning then acts to change the functional form of $\nu_F(\theta;\vec{\omega}_F)$ and $d_S(\theta;\vec{\omega}_S)$ by modifying $\vec{\omega}$. The second variant, shown in SI Figs S5 and S7, considers a "prestructured" policy in which the heading-dependence in both $\nu_F(\theta;\vec{\omega})$ and $d_S(\theta;\vec{\omega})$ is structured with respect to a "goal heading" θ_G. In this case, the flexible parameters $\vec{\omega} = \theta_G$ specify the goal heading, and learning acts by shifting the goal heading while preserving the structured heading dependence in $\nu_F(\theta;\vec{\omega})$ and $d_S(\theta;\vec{\omega})$. The final variant, shown in Figs 4-5, considers a circuit-based implementation of learning that is informed by physiology and connectomic data (see *SI: Linking the Conceptual Model to Known Anatomy*).

In what follows, we outline the assumptions and training algorithms for these model variants. In each variant, we remove the temporal variability in the duration of saccades by assuming that $P(\Delta t|\dot{\theta},\theta,S) = \delta(\Delta t - t_S)$, where t_S is a constant (we will assume $t_S = 300$ms, based on the analysis described in *Analysis Methods: Characterizing saccade properties*; see Table 1).

Flexible policy for maintaining a behavioral preference.

Policy. We approximate the limited angular resolution of the heading representation via a set of radial basis functions $g_i(\theta)$ $(i = 1...n)$ that mimic the anatomical tiling of compass neurons in the Ellipsoid Body. Unless otherwise specified, we used $n = 16$ von Mises basis functions that uniformly tiled the range $[0, 360]$, with concentration factor $\kappa = 8$. With this representation, heading dependence in the drift rate of fixations ν_F and the directional bias of saccades d_S can be achieved by constructing different weighted combinations of these basis functions, with weights $\vec{\omega}_F$ controlling the heading dependence in ν_F, and weights $\vec{\omega}_S$ controlling heading dependence in d_S:

$$\begin{aligned}\nu_F(\theta;\vec{\omega}_F) &= f(\vec{\omega}_F^T\vec{g}(\theta);k_F,f_{0F},f_{MF}) \\ d_S(\theta;\vec{\omega}_S) &= f(\vec{\omega}_S^T\vec{g}(\theta);k_S,f_{0S},f_{MS})\end{aligned} \tag{17}$$

where $\vec{\omega}_a^T\vec{g}(\theta)$ $(a \in \{F,S\})$ is a weighted sum over basis functions evaluated at θ. The sigmoidal function $f(x;k,f_0,f_M) = f_M/(1+\exp(-kx)) - f_0$ enforces bounds on the drift rate of fixations and the probability of a rightward saccade, given a set of parameters $[k,f_0,f_M]$ that control the slope, minimum, and maximum values of the sigmoid. We chose the values of these parameters to bound the drift rate between 0.01 and 1.01, and to bound the probability of rightward saccades between 0 and 1.

As noted above, this policy depends implicitly on a set of fixed parameters $\vec{\beta}$, which now includes the additional parameters that specify this flexible model: $\vec{\beta} = [\delta t, a_F, \eta_F, \varphi_S, \sigma_S, a_S, \eta_S, n, \kappa, k_F, f_{0F}, f_{MF}, k_S, f_{0S}, f_{MS}]$. The values of these parameters are listed in Table 1.

Training. The policy gradients are given by:

$$\begin{aligned} \nabla_{\vec{\omega}} \nu_F(\theta; \vec{\omega}) &= f'(\vec{\omega}_F^T \vec{g}(\theta)) \vec{g}(\theta) \\ \nabla_{\vec{\omega}} d_S(\theta; \vec{\omega}) &= f'(\vec{\omega}_S^T \vec{g}(\theta)) \vec{g}(\theta) \end{aligned} \tag{18}$$

where $f'(x) = df/dx$ is the derivative of the sigmoidal function f. Together with Eqs. (13), (14)-(16), these gradients can be used to compute the update to $\vec{\omega}_F$ and $\vec{\omega}_S$. Training thus acts to change the heading dependence in fixations and saccades by reweighting the basis functions through changes in $\vec{\omega}_F$ and $\vec{\omega}_S$, as detailed in Algorithms 1-2.

In the main text, we used this flexible policy to determine how the average duration of fixations and direction of saccades should be structured as a function of heading in order to (i) maintain a preference for a specific visual pattern P_G (with orientation $\theta(P_G)$), or (ii) maintain a goal heading θ_G. In the first case, we used $n = 16$ von Mises basis functions that uniformly tiled the range $[0, 180]$. The weights corresponding to these basis function specify the actions that agent takes with respect to a particular orientation of a visual pattern; thus, these weights specify the same actions for the two possible orientations of the visual scene that produce the same orientation of a given visual pattern. In the second case, we used $n = 16$ von Mises basis functions that uniformly tiled the range $[0, 360]$.

We used the following sensory response function that decays linearly with angular distance from the preferred heading:

$$R(\theta) = \begin{cases} \frac{\pi}{10}\left(\frac{1}{2} - \min(|\theta - \theta(P_G)|)\right) & \text{preference for goal pattern } P_G \\ \frac{\pi}{10}\left(\frac{1}{2} - |\theta - \theta_G|\right) & \text{preference for goal heading } \theta_G \end{cases} \tag{19}$$

where $\min(|\theta - \theta(P_G)|$ returns the distance to the nearest orientation of a preferred pattern.

Algorithm 1: Learn flexible policy parameters $\vec{\omega}$ via policy-gradient method

input: parametrized policy $\pi(\dot{\theta}, \Delta t | \theta, \xi; \vec{\omega})$
define: total simulation time T_{tot}; fixed policy parameters $\vec{\beta}$
initialize: policy parameters $\vec{\omega} \in \mathbb{R}^d$; integrator $\xi = 0$; time $t = 0$; heading $\theta \in [0, 360]$

while $t < T_{tot}$ **do**
  sample action from policy
  $\dot{\theta}, \Delta t, \Delta\xi \sim \pi(\cdot | \theta, \xi; \vec{\omega})$
  observe sensory response
  $r \leftarrow R(\theta + \dot{\theta}\Delta t)$
  update policy parameters
  $\vec{\omega} \leftarrow \vec{\omega} + r \nabla_{\vec{\omega}} \log \pi(\dot{\theta}, \Delta t | \theta, \xi; \vec{\omega})$
  update heading, time, integrator
  $\theta \leftarrow \theta + \dot{\theta}\Delta t$
  $t \leftarrow t + \Delta t$
  $\xi \leftarrow \xi + \Delta\xi$
end while
return $\vec{\omega}$

Algorithm 2: Sample action $\dot{\theta}, \Delta t$ from flexible policy $\pi(\dot{\theta}, \Delta t|\theta, \xi; \vec{\omega})$

inputs: heading θ; integrator ξ; flexible policy parameters $\vec{\omega}$; fixed policy parameters $\vec{\beta}$; basis functions $\vec{g}(\theta)$

get current drift rate, directional bias
$\nu_F(\theta; \vec{\omega}_F) \leftarrow f(\vec{\omega}_F^T \vec{g}(\theta); k_F, f_{0,F}, f_{M,F})$
$d_S(\theta; \vec{\omega}_S) \leftarrow f(\vec{\omega}_S^T \vec{g}(\theta); k_S, f_{0,S}, f_{M,S})$
integrate drift signal
$\Delta\xi \sim \mathcal{N}(\nu_F(\theta; \vec{\omega}_F)\delta t, \eta_F^2 \delta t)$
if $\xi + \Delta\xi > a_F$ **then**
  saccade
  if $\mathrm{rand}(\cdot) < d_S(\theta; \vec{\omega}_S)$ **then**
    $\dot{\theta} \sim +\mathrm{logn}(\varphi_S, \sigma_S^2)$ (CW turn)
  else
    $\dot{\theta} \sim -\mathrm{logn}(\varphi_S, \sigma_S^2)$ (CCW turn)
  end if
  $\Delta t \leftarrow t_S$
else
  fixate
  $\dot{\theta} \leftarrow 0$
  $\Delta t \leftarrow \delta t$
end if
return $\dot{\theta}$,$\Delta t, \Delta\xi$

Prestructured policy tethered to a single goal heading.

In Fig 2i, we showed how a prestructured goal-heading-dependent policy (derived via the learning algorithm described in the previous section), when tethered to an unstable internal representation of heading, could qualitatively capture the observed structure of fly behavior. Here, we illustrate a simplified form of this prestructured policy, and we show how learning could act to shift the goal heading based on experience.

Policy. In what follows, we will use θ_A to specify the angular orientation of the fly in arena coordinates, and we will distinguish this from the orientation θ_C of the compass heading bump. Before accounting for instabilities in the heading bump, we assume that changes in the arena heading $\Delta\theta_A$ are accompanied by the same change in bump heading $\Delta\theta_C$, such that $\Delta\theta_A = \Delta\theta_C$ and $\dot{\theta}_A = \dot{\theta}_C = \dot{\theta}$ (in the follow section, we will explicitly account for the fact that in the fly heading circuit, the bump heading and the arena heading move in opposite directions during saccades). Bump jumps, which arise from symmetries in the visual environment, will further alter the bump heading θ_C relative to the arena heading θ_A.

The prestructured policy specifies the heading dependence in the drift rate of fixations ν_F, and the directional bias of saccades d_S, relative to a goal heading $\theta_G \in [0, 360)$. Based on the results shown in Fig 2f-g, and given the prevalence of sinusoidal signals in the central complex, we define these dependencies to have the following functional forms:

$$\begin{aligned} d_S(\theta_C;\theta_G) &= -\frac{G_S}{2}\sin(\theta_C - \theta_G) + B_S + \frac{1}{2} \\ \nu_F(\theta_C;\theta_G) &= \frac{G_F}{2}(1 - \cos(\theta_C - \theta_G)) + B_F \end{aligned} \tag{20}$$

where B_S and G_S control the baseline direction and heading-dependence of saccades, and B_F and G_F similarly control the baseline duration and heading dependence of fixations. As before, the average duration of fixations at any given heading θ_C is given by $a_F/\nu_F(\theta_C;\theta_G)$, where a_F is the threshold of the drift diffusion process.

We showed in the main text that the heading representation is unstable in symmetric scenes, and that this instability is manifested in bump jumps whose size reflects the degree of symmetry. We consider the two-fold symmetric scene used in the main text, for which a bump jump results in an angular change of $\Delta\theta_J = \pm 180°$. We further assume that the probability of a jump varies nonuniformly with θ_C. We expect this probability to scale inversely with residency time, such that the bump has the lowest probability of jumping at a location where the residency time has been high, and vice versa. Because the goal heading defines the angular location of highest residency, we define the probability of a jump $p_J(\theta_C;\theta_G)$ to be sinusoidal, with a value that depends on the angular distance from the goal heading:

$$p_J(\theta_C;\theta_G) = \frac{G_J}{2}(1 + \cos(\theta_C - \theta_G)) + B_J \tag{21}$$

where B_J and G_J control the baseline probability and heading-dependence of bump jumps. When G_J is 0, the bump has the same probability of jumping at every heading, determined by the value of B_J. For G_J larger than 0, the probability of a jump will vary nonuniformly with θ, and the minimum probability will be determined by B_J. The jump probability determines whether or not the heading will jump by $\Delta\theta_J$:

$$\theta_C = \begin{cases} \theta_C & \text{with probability } 1 - p_J(\theta_C;\theta_G) \\ \theta_C + \Delta\theta_J & \text{with probability } p_J(\theta_C;\theta_G) \end{cases} \tag{22}$$

From the perspective of a given arena heading θ_A, the effective drift rate and saccade probabilities are given by:

$$\begin{aligned} d_S^{\text{eff}}(\theta_A;\theta_G) &= (1 - p_J(\theta_A;\theta_G))d_S(\theta_A;\theta_G) + p_J(\theta_A;\theta_G)d_S(\theta_A + \Delta\theta_J;\theta_G) \\ \nu_F^{\text{eff}}(\theta_A;\theta_G) &= (1 - p_J(\theta_A;\theta_G))\nu_F(\theta_A;\theta_G) + p_J(\theta_A;\theta_G)\nu_F(\theta_A + \Delta\theta_J;\theta_G) \end{aligned} \tag{23}$$

These curves were illustrated in SI Fig S5.

As before, this policy depends implicitly on a set of fixed parameters $\vec{\beta}$, which now includes the additional parameters that specify the structure setpoint policy: $\vec{\beta} = [\delta t, a_F, \eta_F, \varphi_S, \sigma_S, a_S, \eta_S, G_S, G_F, G_J, B_S, B_F, B_J]$. The values of these parameters are listed in Table 1.

Training. The prestructured policy can be trained using the same policy gradient algorithm discussed above. We briefly discuss a simple algorithm for implementing this. We assume that the gain parameters $\{G_S, G_F, G_J\}$ and baseline parameters $\{B_S, B_F, B_J\}$ remain fixed during training, but that the goal heading θ_G is updated during training and only after a saccade. We further assume that bump jumps occur just after a saccades, and that the goal heading is updated only after the bump has had an opportunity to jump. Thus, a single update consists of the following sequence of steps: (i) sample the duration of a fixation from an inverse Gaussian distribution, (ii) sample the direction and size of saccade, (iii) execute the change in heading in both arena and bump coordinates, (iv) flip a biased coin to determine whether the bump will jump; if the bump jumps, shift the bump location by $180°$, and (v) update the location of the goal heading.

Because the goal heading is shifted only after a saccade (and not during the process of fixation), we can further simplify the process by directly sampling the fixation duration from an inverse Gaussian distribution; i.e., a fixation of total duration Δt is sampled directly from $P(\Delta t|\dot{\theta}, \theta_C, F) = \mathrm{IG}(a_F/\nu_F(\theta_C;\theta_G), a_F^2/\eta_F^2)$, rather than being generated in a timepoint-by-timepoint manner. We then sample a saccade of angular velocity $\dot{\theta} \sim \pi_S(\dot{\theta}|\theta_C;\theta_G)$ and fixed duration t_S (where $\pi_S(\dot{\theta}|\theta_C;\theta_G) = P(\dot{\theta}|\theta_C, S;\theta_G)$ is given in Eq. (12)). Finally, we update the location of the goal heading based on the gradient of the policy with respect to goal heading:

$$\Delta\theta_G(\dot{\theta}^*, \theta_A^*, \theta_C^*;\theta_G) = R(\theta_A + \dot{\theta}\Delta t_S)\frac{1}{\pi_S(\dot{\theta}|\theta_C;\theta_G)}\frac{\partial \pi_S(\dot{\theta}|\theta_C;\theta_G)}{\partial \theta_G}\bigg|_{\dot{\theta}^*,\theta_A^*,\theta_C^*} \tag{24}$$

where $\Delta\theta_G(\dot{\theta}^*, \theta_A^*, \theta_C^*;\theta_G)$ specifies the change in the location of the goal heading that is produced by initiating a saccade with velocity $\dot{\theta}^*$ (and fixed duration t_S) from an arena heading θ_A^* and bump heading θ_C^*, given an initial goal heading of θ_G. The gradient is given by:

$$\begin{aligned}\frac{\partial \pi_S}{\partial \theta_G} &= \frac{\partial}{\partial \theta_G}P(\dot{\theta}|\theta_C, S;\theta_G)\\ &= \mathrm{sgn}(\dot{\theta})\,\mathrm{logn}(|\dot{\theta}|;\varphi_S,\sigma_S)\frac{\partial}{\partial \theta_G}d_S(\theta_C;\theta_G)\\ &= -\frac{G_S}{2}\,\mathrm{sgn}(\dot{\theta})\,\mathrm{logn}(|\dot{\theta}|;\varphi_S,\sigma_S)\cos(\theta_C - \theta_G)\end{aligned} \tag{25}$$

Circuit Model

In Figs 4-5, we developed a circuit-based implementation of the prestructured policy discussed in the previous section, and we used Hebbian learning to update the parameters of that model (note that this differs from the policy-gradient algorithms discussed above). We constructed this circuit model from populations of so-called columnar neurons that tile $360°$ of angular space. In what follows, we parametrize this space by θ, and we express all quantities as functions of θ. See *SI: Linking the Conceptual Model to Known Anatomy* for a discussion of the relationships between this model and anatomical and functional observations.

Policy. In the EB, the current heading θ_C is represented by a bump of activity maintained by a population of columnar compass neurons. We approximate this bump of activity to be sinusoidal with fixed amplitude:

$$r_C(\theta, \theta_C) = \frac{1}{2}\left(\cos(\theta - \theta_C) + 1\right) \tag{26}$$

When the fly turns by an angle $\Delta\theta_A$ in the arena, we assume that this is perfectly captured by the heading circuit, such that the current heading shifts by an equal but opposite angle of $\Delta\theta_C = -\Delta\theta_A$.

Flies are known to exhibit variability in the offset between the orientation of the heading bump and the orientation of the visual scene; we assume here that this offset is zero. In visual scenes without repeating patterns, plasticity between ring neurons and compass neurons ensures that this offset remains stable over time by reinforcing a relationship between visual features in the scenes (as conveyed through the ring neuron receptive fields) and the location of the heading bump. However, in scenes with repeating patterns, ring neurons with a given receptive field will respond similarly when the fly is oriented toward different symmetric views of the same scene. This will result in the strengthening of different sets of ring-to-compass-neuron weights that correspond to the same visual patterns, and will lead to instabilities in the heading representation. We incorporate this instability through a set of heading-dependent compass weights $\vec{\omega}_C = \omega_C(\theta)$ that capture the net inhibition from ring neurons onto compass neurons and thereby determine the relative stability of different headings corresponding to symmetric views of the

visual scene. For the visual scenes considered here, there is a two-fold symmetry, such that the fly sees identical views of the same scene at the two orientations θ_C and $\theta_C + \Delta\theta_J$, where $\Delta\theta_J = 180°$. The relative difference in weights between these two orientations determines the probability that the bump will jump:

$$p_J(\theta_C; \vec{\omega}_C) = \frac{1}{4}\left(1 - \frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\omega_C(\theta) + \frac{\omega_C(\theta_C) - \omega_C(\theta_C + \Delta\theta_J)}{2}\right) \tag{27}$$

where $\frac{1}{N}\sum\omega_C(\theta)$ is the average strength of the weight profile. As in Eq. 22, the jump probability determines whether or not the heading will jump by $\Delta\theta_J$.

From the EB, the heading representation travels through the protocerebral bridge to the fan-shaped body (FB). We assume that information about the fly's current bump heading is combined with information about the goal heading in the FB and then used to drive premotor activity in the lateral accessory lobe (LAL). Specifically, we assume that the information about the goal heading is stored in a set of heading-dependent synaptic weights $\vec{\omega}_G = \omega_G(\theta)$ from tangential motor state neurons onto columnar goal neurons. Here, we assume that the motor state neurons are active (with a constant activity of one, i.e. $r_M(\theta) = 1\ \forall_\theta$) whenever the fly is flying. Thus, the activity profile $r_G(\theta)$ of the goal neurons gives a direct readout of the goal weights:

$$\begin{aligned} r_G(\theta; \vec{\omega}_G) &= r_M(\theta)\omega_G(\theta) \\ &= \omega_G(\theta) \end{aligned} \tag{28}$$

As we will detail below, the set of weights $\vec{\omega}_G$ fully determines the properties of the goal heading.

Finally, we consider populations of output neurons that receive goal activity $r_G(\theta)$ and phase-shifted heading activity $r_C(\theta, \theta_C + \vartheta)$ as inputs, and whose summed output depends on the overlap between current and goal heading through a multiplicative operation:

$$r_O(\theta_C, \vartheta; \vec{\omega}_G) = \frac{\sum_\theta r_C(\theta, \theta_C + \vartheta) r_G(\theta; \vec{\omega}_G)}{\sum_\theta r_C(\theta, \theta_C + \vartheta) r_C(\theta, \theta_C + \vartheta)} + B_O \tag{29}$$

where ϑ is a phase shift, and B_O is a baseline shift (described below). The form of the output activity in Eq. 29 ensures that this output activity will be structured sinusoidally as a function of the fly's current bump heading θ_C relative to the circular mean of the goal weights. To see this, note that the numerator of Eq. 29 can be written as:

$$\begin{aligned} \text{num}[r_O(\theta_C, \vartheta; \vec{\omega}_G)] &= \sum_\theta \frac{1}{2}\left(1 + \cos(\theta - (\theta_C + \vartheta))\right)\omega_G(\theta) \\ &= \frac{1}{2}\sum_\theta \omega_G(\theta) + \frac{1}{2}\left(\frac{e^{-i(\theta_C+\vartheta)}}{2}\sum_\theta e^{i\theta}\omega_G(\theta) + \frac{e^{i(\theta_C+\vartheta)}}{2}\sum_\theta e^{-i\theta}\omega_G(\theta)\right) \\ &= \frac{|\omega_G|}{2} + \frac{r_G}{2}\left(\frac{e^{-i(\theta_C+\vartheta-\theta_G)} + e^{i(\theta_C+\vartheta-\theta_G)}}{2}\right) \\ &= \frac{|\omega_G|}{2} + \frac{r_G}{2}\cos(\theta_C + \vartheta - \theta_G) \end{aligned} \tag{30}$$

where $\sum_\theta e^{i\theta}\omega_G(\theta)d\theta \equiv r_G e^{i\theta_G}$ specifies the modulus r_G and angle θ_G of the circular mean of the goal weights, and $|\omega_G| = \sum_\theta \omega_G(\theta)$ specifies the total strength of the goal weights. As we will next show, saccades will drive the bump heading toward θ_G, and fixations will be maintained longer at θ_G, and thus we define θ_G to be the goal heading. The denominator of Eq. 29 is fixed; we will denote this as $D = \sum_\theta r_C(\theta, \theta_C + \vartheta) r_C(\theta, \theta_C + \vartheta)$.

The dynamic range of the output activity determines how strongly the goal drives behavior; the larger the range, the bigger the differential between the behavior at versus away from the goal location. In such cases with a large differential, we describe the behavior as "highly structured". As can be seen in Eq. 30, this range will be maximized when the modulus of the circular mean, r_G, is maximized. Because we constrain the heading and goal weights to lie in the range $[0, 1]$ (more on this below), the weight profile that maximizes r_G is a square wave in which $N/2$ consecutive weights take a value of 0, and the remaining $N/2$ consecutive weights take a value of 1. Below, we describe a learning rule that will drive the weights toward a sinusoidal profile with a single peak, which approximates this square wave profile. Thus, within the constraints of this learning rule, the more strongly sinusoidal the weight profile, the more structured the behavior.

We construct three different output populations, denoted left ('L'), fixation ('F'), and right ('R'), that differ in the phase shift of their heading-tuned inputs [34]:

$$\begin{aligned} r_L(\theta_C;\vec{\omega}_G) &= r_O(\theta_C,\vartheta=+90;\vec{\omega}_G) = \frac{|\omega_G|}{2D} + \frac{r_G}{2D}\cos(\theta_C+90-\theta_G) \\ r_F(\theta_C;\vec{\omega}_G) &= r_O(\theta_C,\vartheta=180;\vec{\omega}_G) = \frac{|\omega_G|}{2D} + \frac{r_G}{2D}\cos(\theta_C+180-\theta_G) + B_C \\ r_R(\theta_C;\vec{\omega}_G) &= r_O(\theta_C,\vartheta=-90;\vec{\omega}_G) = \frac{|\omega_G|}{2D} + \frac{r_G}{2D}\cos(\theta_C-90-\theta_G) \end{aligned} \tag{31}$$

where we chose $B_L = B_R = 0$, $B_C = \nu_{\max} - |\omega_G|/D$, and $\nu_{\max} = 1.1$; as described below, this ensures that gain of fixations will increase with increasing $|\omega_G|$. We further assume, based on known projection patterns [34], that the populations of left and right output neurons project unilaterally to descending neurons that control leftward and rightward turns, respectively, and that the population of center output neurons projects bilaterally to both sets of descending neurons. We assume that the activity of the right and left output neurons thus determines the average directionality of saccades for a given heading θ:

$$\begin{aligned} d_S(\theta_C;\vec{\omega}_G) &= \frac{1}{2}(1 + r_R(\theta_C;\vec{\omega}_G) - r_L(\theta_C;\vec{\omega}_G)) \\ &= \frac{1}{2} + \frac{r_G}{D}\cos(\theta_C-90-\theta_G) \end{aligned} \tag{32}$$

Note that this function will be largest, and will thus drive the highest probability of clockwise saccades (and counterclockwise bump rotations), when the heading is 90° to the right of the goal heading.

We similarly assume that the activity of the center output neurons determines the drift rate (and thereby the average duration) of fixations:

$$\begin{aligned} \nu_F(\theta_C;\vec{\omega}_G) &= r_F(\theta_C;\vec{\omega}_G) \\ &= \nu_{\max} - \frac{|\omega_G|}{2D} + \frac{r_G}{2D}\cos(\theta_C+180-\theta_G) \end{aligned} \tag{33}$$

$$\langle \Delta t_F(\theta_C;\vec{\omega}_G)\rangle = \frac{1}{\nu_F(\theta_C;\vec{\omega}_G)} \tag{34}$$

Thus, the baseline values of saccade and fixation properties (illustrated in SI Fig S7c,e) are given by:

$$\begin{aligned} \min[d_S] &= \frac{1}{2} - \frac{r_G}{D} \\ \min[\nu_F] &= \frac{2D\nu_{\max} - |\omega_G| - r_G}{2D} \\ \min[\langle \Delta t_F\rangle] &= \frac{2D}{2D\nu_{\max} - |\omega_G| + r_G} \end{aligned} \tag{35}$$

and the gain of saccade and fixation properties (illustrated in SI Fig S7b,d) are given by:

$$\begin{aligned} \text{gain}[d_S] &= \frac{2r_G}{D} \\ \text{gain}[\nu_F] &= \frac{r_G}{D} \\ \text{gain}[\langle \Delta t_F\rangle] &= \frac{r_G/D}{\nu_{\max}^2 + |\omega_G|^2/(4D^2) - r_G^2/(4D^2) - \nu_{\max}|\omega_G|/D} \end{aligned} \tag{36}$$

As can be seen from Eqs. 32-34, the modulus r_G of the circular mean of the goal weights determines the gain of both the fixational drift rate and the saccade directionality; the larger r_G, the larger the turn bias when the heading bump is to the right or left of the goal heading, the longer the fixations at the goal heading, and the shorter the fixations away from the goal heading. In this way, the multiplicative operation performed by the output neuron populations (and specified by Eq. 30) guarantees that the fly's internal policy will remain structured as a function of the current heading relative to the goal heading, regardless of their specific values.

Training. We assume that there is plasticity in both $\vec{\omega}_C$ and $\vec{\omega}_G$ that is mediated by the activity of the heading bump $r_C(\theta, \theta_C)$:

$$\Delta\omega_C(\theta, \theta_C; \vec{\omega}_C) = \alpha_C \Delta_C \tag{37}$$

$$\Delta\omega_G(\theta, \theta_C; \vec{\omega}_G) = \alpha_G \Delta_G \tag{38}$$

where

$$\Delta_C = \big[r_C(\theta, \theta_C) - \omega_C(\theta)\big]_+ \Theta(1 - \omega_C) - \big[\omega_C(\theta) - r_C(\theta, \theta_C)\big]_+ \Theta(\omega_C) \tag{39}$$

$$\Delta_G = \begin{cases} +\big[r_C(\theta, \theta_C) - \omega_G(\theta)\big]_+ \Theta(1 - \omega_G) - \big[\omega_G(\theta) - r_C(\theta, \theta_C)\big]_+ \Theta(\omega_G) & R(\theta_A) > 0 \\ -\big[r_C(\theta, \theta_C) - \omega_G(\theta)\big]_+ \Theta(\omega_G) + \big[\omega_G(\theta) - r_C(\theta, \theta_C)\big]_+ \Theta(1 - \omega_G) & R(\theta_A) < 0 \\ 0 & R(\theta_A) = 0 \end{cases} \tag{40}$$

Here, $[\cdot]_+$ denotes rectification, and $\Theta(\cdot)$ is the heaviside function. The first of these plasticity rules is similar to that used in [29] in that the change in weights is proportional to the coactivity between ring neurons (whose activity here is implicilty conveyed through the compass weights) and compass neurons (whose activity here is assumed to have a fixed profile $r_C(\theta, \theta_C)$). We additionally assume that weights are only updated during fixations; in practice, we partitions fixations into time increments of 100ms, and we iteratively update weights at each time increment. The second plasticity rule differs from the first in that it additionally incorporates the valence, $R(\theta_A)$, of the current arena heading θ_A. We assume that this valence is carried by tangential neuromodulatory neurons that innervate the FB and themselves receive input from heading-tuned neurons [34]. We assume this valence takes the following form:

$$R(\theta_A) = \begin{cases} +1 & \theta_A \in \text{safe} \\ -1 & \theta_A \in \text{danger} \end{cases} \tag{41}$$

Algorithms 3-4 detail how this circuit model is implemented and updated through training.

Algorithm 3: Learn heading and goal weights, $\omega_C(\theta)$ and $\omega_G(\theta)$

input: parameterized policy $\pi(\Delta\theta, \Delta t|\theta; \vec{\omega}_C, \vec{\omega}_G)$
define: total simulation time T_{tot}, fixed policy parameters $\vec{\beta}$, learning rates α_C, α_G
initialize: weights $\vec{\omega}_C$ and $\vec{\omega}_G$; bump heading $\theta_C \in [0, 360)$; arena heading $\theta_A = \theta_C$; time $t = 0$, $\Delta t = 0$;

while $t < T_{tot}$ **do**
  sample action from policy
  $[\Delta\theta_S, \Delta\theta_F], [\Delta t_S, \Delta t_F] \sim \pi(\cdot|\theta_C; \vec{\omega}_C, \vec{\omega}_G)$
  update heading and arena heading after saccade
  $\theta_C \leftarrow \theta_C + \Delta\theta_S$
  $\theta_A \leftarrow \theta_A - \Delta\theta_S$
  $t \leftarrow t + \Delta t_S$
  determine whether bump will jump
  if $\mathrm{rand}(\cdot) < p_{\mathrm{jump}}(\theta_C; \vec{\omega}_C)$ **then**
    $\theta_C \leftarrow \theta_C + \Delta\theta_C$
  end if
  observe sensory response
  $r \leftarrow R(\theta_A)$
  update weights $\forall_\theta$
  while $\Delta t < \Delta t_F$ **do**
    $\omega_C(\theta) \leftarrow \omega_C(\theta) + \alpha_C \Delta_C$
    $\omega_G(\theta) \leftarrow \omega_G(\theta) + \alpha_G \Delta_G$
    $\Delta t \leftarrow \Delta t + 0.1$
  end while
  $\Delta t = 0$
end while
return $\vec{\omega}_C, \vec{\omega}_G$

Algorithm 4: Sample action sequence $[\Delta\theta_S, \Delta\theta_F]$, $[\Delta t_S, \Delta t_F]$ from setpoint policy $\pi(\Delta\theta, \Delta t|\theta; \vec{\omega}_C, \vec{\omega}_G)$

inputs: heading θ_C; weights $\vec{\omega}_C$, $\vec{\omega}_G$

get current drift rate and directional bias
$\nu_F(\theta_C; \vec{\omega}_G) \leftarrow r_F(\theta_C; \vec{\omega}_G) + B_F$
$d_S(\theta_C; \vec{\omega}_G) \leftarrow (1 + r_R(\theta_C; \vec{\omega}_G) - r_L(\theta_C, \vec{\omega}_G))/2$
saccade
if $\mathrm{rand}(\cdot) < d_S(\theta_C; \vec{\omega}_G)$ **then**
  $\Delta\theta_S \sim +\mathrm{logn}(\varphi_S, \sigma_S^2)$ (CW)
else
  $\Delta\theta_S \sim -\mathrm{logn}(\varphi_S, \sigma_S^2)$ (CCW)
end if
$\Delta t_S \leftarrow t_S$
fixate
$\Delta\theta_F \leftarrow 0$
$\Delta t_F \leftarrow 1/\nu_F(\theta_C; \vec{\omega}_G)$
return $[\Delta\theta_S, \Delta\theta_F]$, $[\Delta t_S, \Delta t_F]$

General policy parameters			
T_{tot}	240	total simulation time in sec	duration of two trials
t_S	320	duration of saccade in msec	median duration in data
δt	0.001	timescale of drift diffusion process in sec	sampling rate of behavioral data
η_F	1	spread of drift diffusion process	estimated from distribution
a_F	0.79	threshold for drift diffusion process	of fixation durations
φ_S	3.89	parameters of lognormal distribution	estimated from distribution
σ_S	0.54	over saccade sizes (in deg)	of saccade sizes
Parameters for flexible policy			
k_F	1	sensitivity of drift rate	chosen for illustration
$f_{0,F}$	-0.01	sets minimum and maximum scale	constrains avg fixation duration
$f_{M,F}$	10	of drift rate	between 100ms and 120s
k_S	1	sensitivity of saccade probability	chosen for illustration
$f_{0,S}$	-0.01	sets minimum and maximum scale	constrains saccade probability
$f_{M,S}$	0.98	of saccade probability	between 0.01 and 0.99
n	16	number of von Mises functions	matched to EB tiling
κ	8	concentration of von Mises functions	
Parameters for setpoint policy			
G_S	0.9	gain of saccades (controls heading-dependence)	chosen for illustration
B_S	0	baseline direction of saccades	
G_F	0.8	gain of fixations (controls heading-dependence)	chosen for illustration
B_F	0.05	baseline duration of fixations	
G_J	0.7	gain of bump jump (controls heading-dependence)	chosen for illustration
B_J	0.05	baseline probability of jump	
$\Delta\theta_J$	180	size of bump jump in deg	matched to data
Parameters for circuit implementation of setpoint policy			
N	32	discretization of heading space	chosen for illustration
α_C	0.002	learning rate of compass weights	chosen for illustration
α_G	0.001	learning rate of goal weights	

Table 1: Parameter values used in RL and circuit models.