Perceptual learning improves discrimination while distorting appearance

Sarit F.A. Szpiro[1]*, Charlie S. Burlingham[2]*, Eero P. Simoncelli[2,3,4,5], Marisa Carrasco[2,3]

* Joint first authorship

[1] Department of Special Education, University of Haifa, Israel
[2] Department of Psychology, New York University, New York, NY 10003
[3] Center for Neural Science, New York University, New York, NY 10003
[4] Courant Institute of Mathematical Sciences, New York University, New York, NY, USA
[5] Flatiron Institute, Simons Foundation, New York, NY, USA

Significance

Perceptual learning refers to the increase in sensitivity to small stimulus differences (for example, distinguishing two similar perfumes or shades of blue) that arises from training. Another important perceptual dimension is appearance, the subjective sense of stimulus magnitude. It seems intuitive that training-induced improvements in discrimination would be accompanied by more accurate appearance. We used both discrimination and estimation tasks to test this hypothesis, and find that that while training improves discrimination ability, it leads to increases in appearance distortion. To explain this counterintuitive finding, we propose a model of how distortions of appearance can arise from increased precision of neural representations and serve to enhance distinctions between perceptual categories, a potentially important property of real-world perceptual learning.

Abstract

Perceptual learning, a form of adult brain plasticity that produces improved discrimination, has been studied in various tasks and senses. However, it is unknown whether and how this improved discrimination alters stimulus appearance. Here, in addition to a discrimination task, we used an estimation task to investigate how training affects stimulus appearance in human adults. Before and after training, observers were shown stimuli composed of dots moving slightly clockwise or counter-clockwise horizontal, whose appearance has been shown to be biased away from horizontal. Observers were subdivided into three groups: Those who (1) trained in a discrimination task; (2) trained in an estimation task; (3) did not train. Training improved discrimination accuracy and decreased coherence thresholds. Counterintuitively, training also distorted appearance, substantially exacerbating estimation biases. These changes occurred in both training groups (but not in the no-training control group), suggesting a common learning mechanism. We developed a computational observer model that simulates performance on both discrimination and estimation tasks. The model incorporates three components: (1) the internal representation favors cardinal motion directions, which are most common in the natural environment; (2) in the estimation task, observers implicitly categorize motion, conditioning their estimates on this; and (3) both types of training induce an increase in the precision of representation of trained motions. We find that the simulations of the model, fit to individual observer data, can account for their improved discrimination and increased estimation bias. We conclude that perceptual learning improves discrimination while simultaneously distorting appearance.

Introduction

One of the most remarkable forms of adult brain plasticity is the capacity to develop perceptual expertise. Perceptual learning (PL) is defined as long-lasting improvements in the performance of perceptual tasks following practice (for reviews, see (1–3)). PL has been documented in every sensory modality (4–6), and has been extensively studied using detection and discrimination tasks in behavior (e.g., (7–12), electrophysiology (5, 13–19), neuroimaging (20–24) and computational models (25–28). Explanations of the mechanisms underlying PL range from low-level changes in sensory neurons — e.g., changes in tuning curve shape (13, 14) and gain (16, 24), feed-forward weights, and noise correlations (27) — to top-down modulations of neural responses based on context and task (17–19), changes in decision-making (5), and reweighting (12, 28).

Most studies of perceptual learning have focused on task performance (e.g., improved discrimination or detection accuracy for a given stimulus attribute). However, another critical aspect of perception is **appearance**, the subjective perceived value of a stimulus attribute, which can be assessed via estimation tasks. Appearance can be examined in terms of both precision (the variability over repeated trials) and bias (the mismatch between the average percept and the true stimulus value). Previous studies suggest that PL improves precision, reducing thresholds in detection and discrimination tasks (29–33). Yet, whether and how PL affects bias remains unknown. On the one hand, the improved precision could be accompanied by reduced bias (a more veridical appearance). Alternatively, improved precision could come at the expense of an increase in bias (a more distorted appearance). The effect of PL on perceptual bias has implications for theories and computational models (12, 34, 35), neurophysiology (36), and philosophy (37) of learning and perception.

A well-established example of perceptual bias is “reference repulsion”, in which estimation of orientations and motion directions are biased away from cardinal (horizontal and vertical) directions (38–41). The only PL study to date that has investigated appearance revealed that training (without feedback) to estimate nearly horizontal motion *increased* repulsive biases away from horizontal, as measured through either an explicit estimation report of motion direction, or the direction of smooth pursuit eye movements (8). The mechanism behind this surprising result is unknown. Furthermore, these results raise the question of how training on a discrimination task — as has been used in nearly all PL studies — affects estimation biases. If training on a discrimination task not only improves accuracy but also exacerbates repulsive biases, these findings would raise fundamental questions about how PL affects perceptual representations to improve performance, and the mechanisms that underlie this improvement.

Here, we investigated discriminability and appearance by measuring the behavioral consequences of training on both discrimination and appearance. Observers were initially tested (“pre-training”) on both a discrimination task (indicate “clockwise” or “counterclockwise”) and an estimation task (indicate perceived motion direction by adjusting the orientation of a line) (**Fig. 1A**). All observers exhibited substantial repulsive biases in the estimation task (e.g., for a 4º motion stimulus they estimated a 12º response), consistent with previous reports of reference repulsion (38, 39). Sometimes observers made estimates that were on the other side of the horizontal boundary (i.e., for a +4º stimulus, they estimated a -12º response, “a misclassified estimate”). After three days of training on either task, observers exhibited improved discrimination accuracy and decreased noise coherence thresholds to a similar degree. Strikingly, training also exacerbated the estimation biases (e.g., after training, a 4º

motion stimulus was estimated as 18º), while reducing the frequency of misclassified estimates. Thus, PL improved performance at the expense of a less accurate representation of the presented stimulus.

To explain how PL exacerbated estimation bias while improving discrimination, we developed a computational observer model that makes predictions of discrimination accuracy and the distribution of estimation judgments before and after learning. The model combines three main components, which have been suggested in the literature in other contexts: (1) Motion directions near horizontal are over-represented by the brain (42), leading to perceptual biases away from the horizontal boundary even before learning; (2) Learning increases the precision of the internal representation of motion (29–33); and (3) When performing the estimation task, the observer first (implicitly) categorizes the stimulus motion direction relative to horizontal, and discards evidence inconsistent with this choice. This decision strategy, known as 'conditional inference' or ‘confirmation bias’, has previously been used to explain estimation biases as well as ‘misclassified estimates’ (34, 43). We fitted our model to data of individual observers and found that the combination of these three components accounts for the observed improvements in discrimination, as well as in the full distribution of estimation responses, for each observer.

Results

To examine how PL modifies both discrimination ability and appearance, we asked observers to make judgments about random dot stimuli moving in near-horizontal directions. Observers indicated their responses in either a discrimination task (specifying “clockwise” or “counter-clockwise” relative to rightward horizontal), or an estimation task (indicating the perceived direction of motion by adjusting the orientation of a line). Initially, all observers performed both tasks for motion directions of ±2°, ±4° and ±8° relative to a horizontal motion direction (**Fig. 1A-B**). Following this pre-training, one group (n=7) was trained on the ±4° estimation task, and another group (n=7) was trained on the ±4° discrimination task, for three consecutive days. A third group, the control group (n=7), performed the pre-training tests, but received no additional training over the next three days. Finally, observers from all three groups were tested post-training on both tasks. No feedback was provided at any time (see Discussion).

All stimuli consisted of a mixture of dots moving coherently in a common direction and a subset moving randomly. Before the experiment, each observer’s noise coherence threshold (the fraction of coherent dots yielding 75% discrimination accuracy) was obtained for the ±4° motion directions using a staircase procedure. This threshold was subsequently used in testing and training sessions, and the coherence threshold was assessed again after the posttest was completed. In the pretest, neither discrimination accuracy nor estimation judgments interacted with training group for either stimulus motion direction (both $F(4,36)<1$), demonstrating that prior to training the groups had similar performance in both tasks.

PL improved discrimination accuracy

For both training groups, PL improved discrimination of motion direction (**Fig. 1C,E (bottom panel)**, **Fig. S1**), consistent with prior studies (7, 44). A mixed-design ANOVA revealed a 2-way significant interaction (Training groups vs Control group X Session: $F(1,19)=5.79$, $p=0.02$): discrimination

accuracy improved significantly between the pre-training and the post-training sessions for observers who trained (Session: $F(1,12)=17.35$, $p=.001$), and to a similar degree in both training tasks (Training Task x Session: $F(1,12)<1$) and across directions (Training Task X Session X Direction: $F(2,24)<1$), but accuracy was unchanged for observers who did not train (Control group: $F(1,6)<1$). In sum, discrimination improved regardless of training task and to a similar extent for the trained and untrained motion directions.

Noise thresholds decreased after training

We analyzed noise coherence thresholds before and after training to determine whether there was an improvement in signal-to-noise, as has been found in PL research (e.g., (29–33). There was a significant 2-way interaction between group and session[1] (Training groups vs Control group X Session: $F(1,16)=8.9$, $p=0.009$; **Fig. 1D**). In the control group, without training, coherence thresholds did not change significantly between pretest and posttest sessions, from 47.3% to 44.9%, $t(4)=0.04$, $p=0.99$). In contrast, training significantly reduced coherence threshold to a similar degree for both training groups (*Session*: $F(1,10)=22.64$, $p<0.001$; *Training Task X Session*: $F(1,10)<1$): For estimation-training from 48.8% to 34.3% ($t(6)=3.55$, $p=0.01$); for discrimination-training, from 47.1% to 33.1% ($t(5)=2.97$, $p=0.03$, **Fig. 1D**). Across observers, changes in noise coherence thresholds between posttest and pretest were correlated with changes in discrimination accuracy between posttest and pretest ($r=-0.66$, $p<0.001$).

Estimates were repelled from horizontal before training

Previous studies have demonstrated substantial repulsive biases in perceived motion directions near the cardinal axes (8, 38, 39). Consistent with this repulsive bias, our observers estimated motion directions near horizontal as being further away from the horizontal boundary (**Fig. 1F**): On average, ±2° motion was estimated as ≈±4.6°, ±4° as ≈±11.8°, and ±8° as ≈±19.1°. In addition, the distributions of estimates were often distributed bimodally, with a primary mode on the 'correct' side of the boundary (i.e., corresponding to the true category of the stimulus, e.g., "CW"), but centered at a biased value (e.g., 15.1° for observer O4 and directions both ±4° collapsed such that estimates with a positive sign correspond to the correct category**;** **Fig. 2**), and a secondary mode on the 'incorrect' side of the boundary (e.g., "CCW"), centered at approximately the negative of the primary mode (e.g., -9.5° for observer O4 and directions ±4° collapsed such that estimates with a negative sign correspond to the incorrect category; **Fig. 2**). In a few cases, there was a third mode centered near the horizontal boundary (**Fig. S1**).

We quantified the shape of these distributions by fitting each with a Gaussian Mixture Model containing between one and three components. In the pretest, the majority of the estimation distributions were bimodal (63%, over all observers and directions), and the remainder were either unimodal (24%), or tri-modal (13%) (Supplementary Materials, **Fig. S1**). Bimodal results have been obtained in experiments that require an explicit categorization judgement prior to the estimation response (34, 43, 45, 46). As in these studies, we hypothesized that observers performed an implicit categorization that influenced their subsequent estimation. Were that the case, we would expect that conditions with more errors in the discrimination task (i.e., subject reports "up" when the right answer is "down" or vice versa), would also be associated with a larger fraction of 'incorrect' estimations (i.e.,

[1] Due to a technical problem, we do not have the noise thresholds at posttest for three observers, we focus our analysis on the remaining 18 participants.

subject adjusts the line clockwise when the true direction was counter-clockwise, relative to horizontal or vice versa). Thus, we predicted that discrimination accuracy should be correlated with the categorization accuracy of estimation judgements, that is the proportion of estimates that fall within the true stimulus category (e.g., "CW" or "CCW"), which here we call "***signed estimation accuracy***." Indeed, we found a strong correlation between signed estimation accuracy and discrimination accuracy, collapsed across motion direction, training group for both testing sessions (overall: $r=0.66$, $p<1x10^{-6}$; pretest: $r=0.84$; $p<1x10^{-6}$; posttest: $r=0.82$, $p<1x10^{-6}$, **Fig. 1G**), and no significant difference between sessions (Williams test: $p=0.766$). These results, along with the bi-modal shape of the distribution, support the hypothesis that observers performed an implicit categorization in all tasks and conditions, even when they were not instructed to do so.

Perceptual learning increased overestimation

Training improved discrimination accuracy but increased estimation biases. We first summarized performance in the estimation task with the average of the estimate distribution. There was a significant 2-way interaction in the average estimate between observers in the training and control groups (Training vs Control groups X Session: $F(1,19)=5.86$, $p=0.02$). In the control group, average estimates did not change (Session: $F(1,6)<1$). Remarkably, in the training groups there was a substantial *increase* in overestimations (Session: $F(1,12)=29.4$, $p<0.001$). For example, directions of ±4°, which were perceived as ≈±12° in the pretest, were perceived as ≈±17.85° in the posttest (**Fig. 1F**). Overestimation increased to a similar extent across the three tested motion directions (Session X Direction: $F(2,24)=2.06$, $p=0.14$): on average, an increase of 6°±0.33°. The increase in overestimations between pretest and posttest was correlated with the decrease of noise coherence thresholds across observers ($R=-0.49$, $p<0.01$).

For both training tasks, overestimation increased to a similar extent for trained and untrained motion directions (Training Task X Session X Direction: $F(2,24)<1$). There was a significant 2-way interaction (Training task X Session: $F(1,12)=4.86$, $p=0.04$): for both training groups the Session was significant, with a stronger effect in the Estimation ($F(1,6)=17.9$, $p<.005$) than in the Discrimination ($F(1,6)=13.8$, $p=.01$) training groups. In sum, for both training groups, training modifies appearance by increasing overestimation away from the horizontal for both the trained and untrained directions.

The learning-induced change in the mean of the estimate distribution could be the result of a variety of different changes in the underlying distribution. For example, the mode of the sub-distribution on the correct side of the discrimination boundary could shift away from the boundary, or the proportion of incorrectly classified estimates could be reduced (i.e., the observer makes fewer mis-categorizations, and signed estimation accuracy increases (**Fig. S2**)). To distinguish these two possibilities, we examined the primary mode of the estimate distribution. Unlike the mean, we found that the mode was not significantly altered by training (Training groups vs. Control group X Session: $F(1,19)<1$), although we did note a small but detectable mode shift for a subset of participants (**Fig. S1, Fig. S3**). On the other hand, training increased signed estimation accuracy in both training groups, regardless of training task and across directions (*Session*: $F(1,12)=25.9$, $p<0.001$; *Session X Direction X Training Task*: $F(2,24)<1$). In contrast, in the control group signed estimation accuracy (proportion of estimates in the correct category) did not significantly change (*Session*: $F(1,6)= 3.45$, $p=0.112$; *Session X Direction*: $F(2,12)<1$). Thus, for most observers who trained, PL modified the estimation mean primarily by improving signed estimation accuracy. And for some of these, PL additionally increased the bias in the mode of the estimation (**Fig. 2C; Fig. S3**).

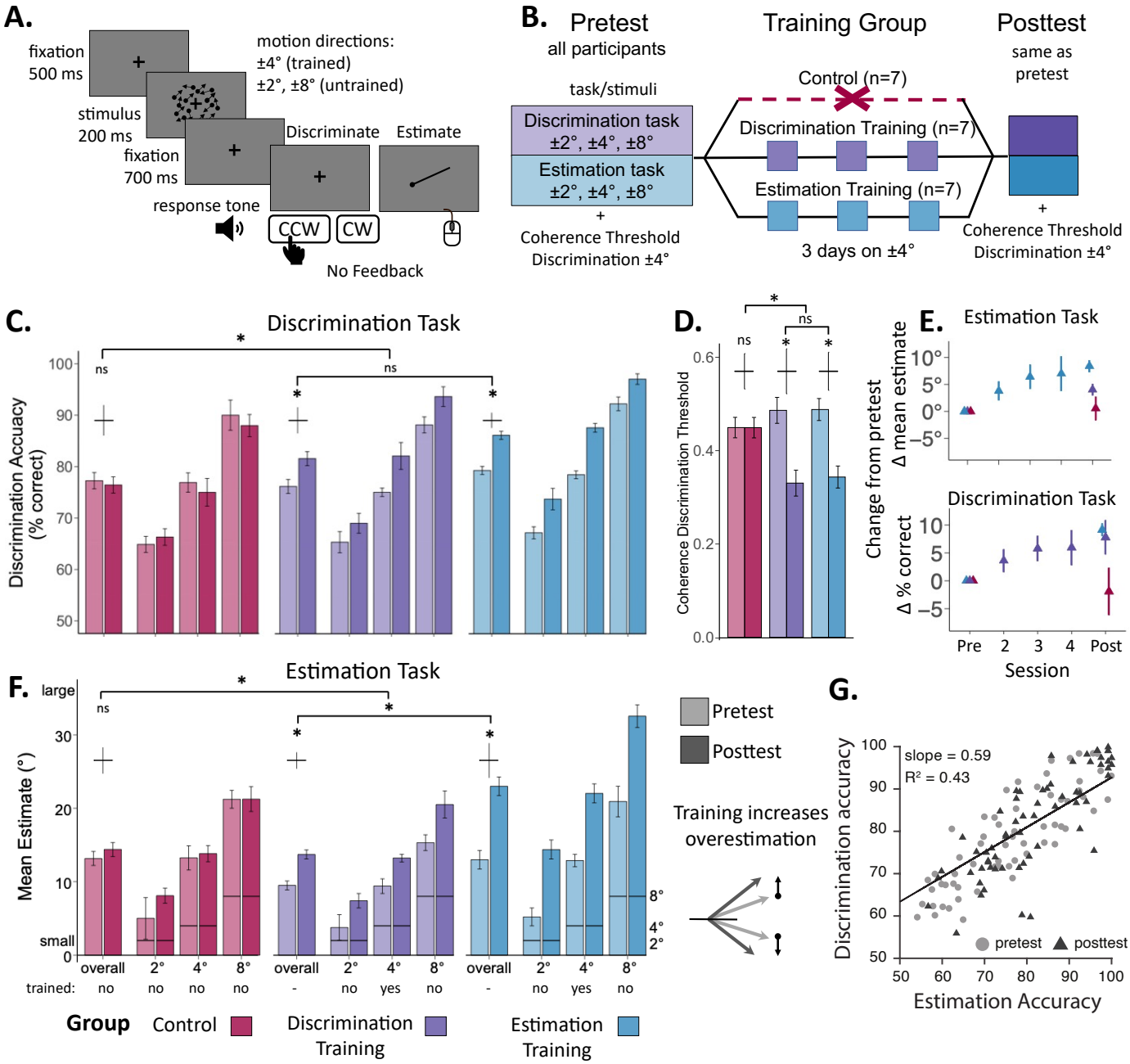

Fig. 1. Experimental protocol and behavioral evidence of PL.

A. Trial sequence. **B.** Training procedure for three subject groups. **C.** Average discrimination accuracy across groups (indicated with three colors) at pre and posttest (unsaturated vs saturated colors, respectively) for trained (±4°), untrained (±2° and ±8°), and average across directions. Errors bars represent standard error of the mean with Morey's correction. **D.** Coherence discrimination threshold across groups (colors), pre and posttest; training significantly reduces coherence thresholds. **E.** Performance across days in the estimation task (top), and for the discrimination task (bottom), for the different training groups. Errors bars represent standard error of the mean. **F.** Mean estimates across groups at pre and posttest for trained (±4°), untrained (±2° and ±8°), and average across directions. Errors bars represent standard error of the mean with Morey's correction. Horizontal lines indicate the true stimulus motion directions, highlighting the extent of horizontal repulsion. Performance was significantly modified by training (on both estimation and discrimination): from pretest to posttest,

discrimination accuracy improved, and estimation biases increased. This was not the case in the control group. **G.** Discrimination accuracy and signed estimation accuracy (percent of direction estimates that were consistent with the correct up/down classification) across all observers in pretest (light circles) and posttest (dark triangles). There was a significant correlation across observers and sessions (see text), suggesting that observers implicitly classified motion as clockwise or counterclockwise prior to estimating the direction.

Observer Model

How does PL simultaneously improve discrimination accuracy and worsen estimation biases? To understand these intriguing findings, we developed an observer model that performs and learns both tasks. The observer model performs inference based on a probabilistic neural population code (45, 49). In the encoding stage, a stimulus elicits spikes in a population of motion-direction-sensitive neurons with independent and identically distributed (i.i.d.) Poisson noise, yielding a noisy population response on each trial. The neurons' tuning curves tile the space of motion directions, but are warped such that they over-represent directions near horizontal (**Fig. 2A,** (47, 50)), as has been observed in physiological recordings (42). This type of inhomogeneity is generally consistent with theories of coding efficiency: more neural resources are devoted to the cardinal directions because they are more commonly encountered in the environment (48). We assume that the decoding stage is unaware of this warping (51), and assumes the encoding population is equi-spaced, which leads to perceptual biases away from the horizontal boundary (i.e., boundary avoidance). Given that behavioral performance was similar in the two training groups, we assumed that PL arose from an increase in the precision of the internal representation of motion direction. We modeled this with a gain increase of the sensory neurons encoding the training stimuli (±4º).

The observer model makes either a discrimination or estimation judgement, depending on the task. For discrimination, the observer model reports "counter-clockwise" if the majority of the internal evidence (likelihood function) lies counter-clockwise of the discrimination boundary, and "clockwise" otherwise. For estimation, the model performs an implicit categorization (using the same rule), discards internal evidence inconsistent with the chosen category, and reports the mean of the remaining evidence as its estimate ((35); see Methods for implementation details). We also assume that the behavioral response itself is noisy in both the estimation task (i.e., additive Gaussian motor noise) and discrimination task (i.e., lapse rate).

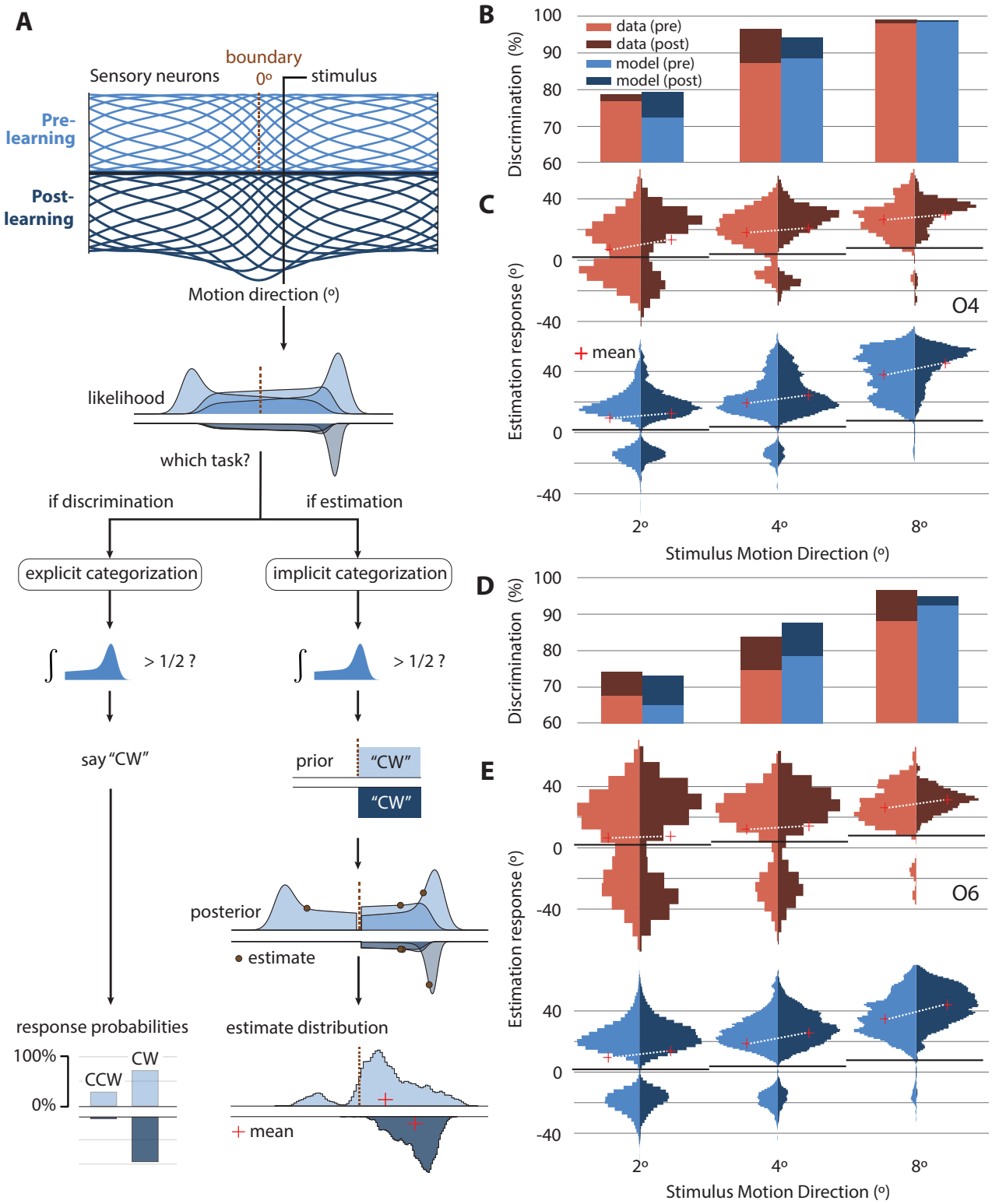

Fig. 2. Observer model and human/model behavior for two observers.

(A) A population of tuned sensory neurons encodes motion direction (tuning curves, representing mean stimulus response**;** Before training - light blue, upper panels; After training – dark blue, inverted, lower panels). The population is warped such that more neurons represent near-horizontal motion directions, even before training (47, 48). Training causes an increase in the gain of the neurons encoding the trained stimuli (±4°), rescaling

their tuning curves (plotted upside down, dark blue). For each trial, spike counts for each cell are drawn from independent Poisson distributions, with firing rate governed by a tuning curve's value for the stimulus's motion direction. The decoder computes the likelihood of different motion directions given these noisy responses and is assumed to be "unaware" of the fact that motion space is warped. This leads to skewed likelihood functions (3 examples shown). Due to sensory noise, likelihoods fluctuate across trials, with modes sometimes falling on the opposite side of the boundary relative to the stimulus. Regardless of task, the observer performs a discrimination judgement, by comparing the mass of the likelihood on the two sides of the boundary. For the discrimination task, this answer is reported. For the estimation task, this "implicit" discrimination judgement conditions the upcoming estimate (34) — specifically, estimates (3 examples indicated with black points) are computed as the mean of the portion of the likelihood on the side of the boundary corresponding to the chosen category. The mean estimate (red cross) is biased away from the true stimulus by both the boundary avoidance (warping) and conditional judgement. And when, by chance, the likelihood falls on the incorrect side of the boundary, its corresponding estimate contributes to a second mode in the estimate distribution, and we label it "misclassified." The training-induced increase in sensory gain generates fewer likelihoods on the wrong side of the boundary and interacts with the warping/unaware decoding to reduce internal evidence for motion near the boundary (i.e., "boundary avoidance"). This reduces misclassifications of motion directions (i.e., increases discrimination accuracy) and increases estimation bias. **(B,D)** Comparison of human and model behavior in the discrimination task for two representative observers in the estimation training group (O4, O6). Bar height, discrimination accuracy (% correct) across trials (N =120). **(C,E)** Comparison of human and model behavior in the estimation task. Histograms of estimates (N = 60). Red cross, mean of estimate distribution. White line, change in mean with learning. Black line, stimulus motion direction. Note bimodal shape of estimate distribution (arising from implicit discrimination judgement) and its dependency on learning and stimulus motion direction.

Model behavior reproduces increased discrimination accuracy and overestimation after PL

For each observer, we fit the model to all behavioral data simultaneously: pre- and post-learning discrimination accuracies and full estimate distributions for all stimuli (±2, ±4 & ±8º). This process yielded parameter estimates shared between the two tasks as well as model behavior for each task separately (**Fig. 1B-D**; also see Methods: Parameter estimation).

The model behavior reproduced many aspects of observer behavior, including training-induced increases in discrimination accuracy (**Fig. 3A**), overestimation (**Fig. 3B**), and signed estimation accuracy (**Fig. 3C**), as well as a tight correlation between discrimination accuracy and signed estimation accuracy before and after learning (**Fig. 1G; Fig. S3B)**). Furthermore, the model quantitatively reproduced the transfer of performance from the trained stimuli (±4º) to the tested stimuli (±2 & ±8º) (**Fig. 2B,C**, **Fig. 3**), including scaling of discrimination accuracy (**Fig. 3A**) and reductions in the bimodality of the estimate distribution with learning (**Fig. 2C**). Specifically, after learning, the model produced fewer estimates that were inconsistent with the true category of the stimulus, leading to a more unimodal estimate distribution, which matched the human observers' estimate distributions. Across observers, the model explained 85% of the variance in the discrimination accuracy data ($p<1x10^{-6}$), 44% of the variance in estimation mean ($p<1x10^{-6}$), and 52% of the variance in signed estimation accuracy ($p<1x10^{-6}$; **Fig. 3**).

The model captured large individual differences that were apparent even before training, including nearly unbiased estimation judgements for some observers (e.g., **Fig. S3**, O12, O15), highly biased estimation judgements for others (e.g., **Fig. 2:** O4, O6**; Fig. S3**: O7), which corresponds to varying degrees of warping in the model (**Fig. S10**), as well as low discrimination accuracy (e.g., **Fig. S3:** O7,

O12) and high discrimination accuracy (e.g., **Fig. 2:** O4; **Fig. S3**: O13). For observers with little to no boundary avoidance, the model behavior was similar to existing models of conditional inference (34, 43). The model reproduced the variability in the estimates in the control group (**Fig. S4**) as well as the clear signs of the implicit categorization decision strategy (i.e., large overestimation pre- and post-training and a tight correlation between estimation and discrimination accuracy).

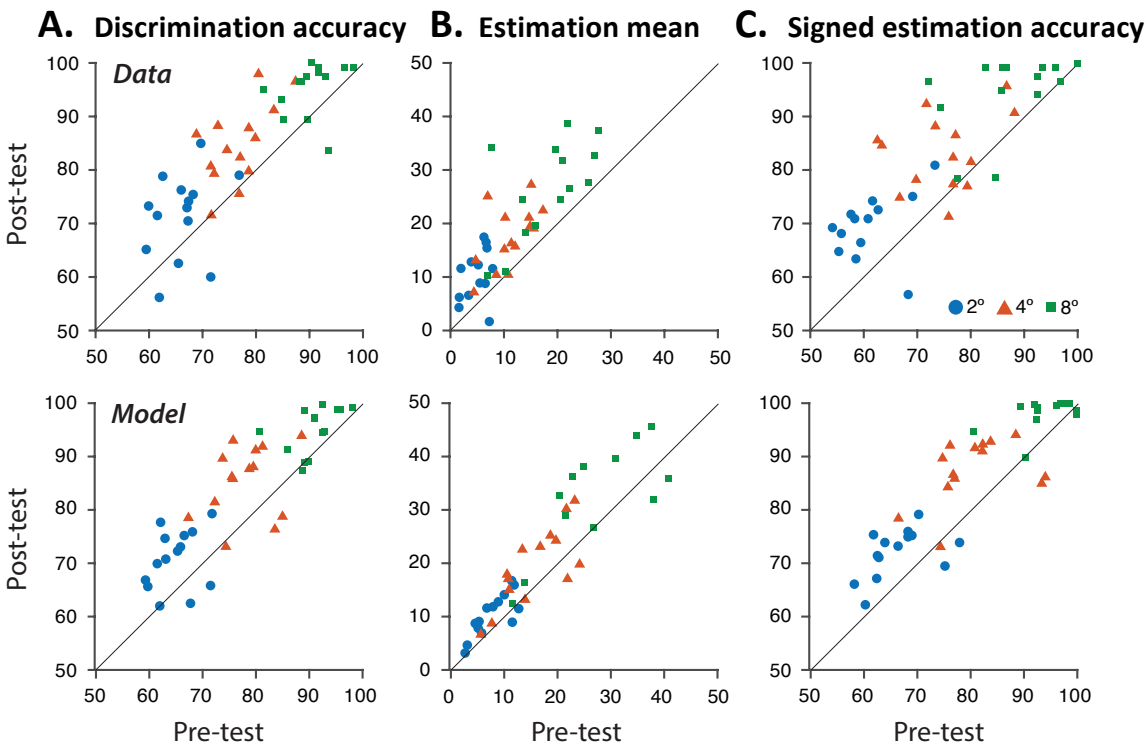

Fig. 3. Comparison of human and model performance, over observers trained on both tasks.

(A) Pre- vs. post-test discrimination accuracy. Top row, human behavioral data. Bottom row, model behavior. Each symbol indicates average discrimination accuracy for one observer for the 2°, 4°, or 8° stimuli (blue circles, red triangles, green squares, respectively). **(B)** Pre- vs. post-test estimation mean. **(C)** Pre- vs. post-test signed estimation accuracy (% of estimates falling on same side of discrimination boundary as the stimulus motion direction). Note similarity between discrimination and signed estimation accuracy (panels A and C) in both data and model.

Importance of model components in explaining human behavior

We used model comparisons to qualitatively assess the causal contribution of each model component to the overall behavior (see **Supplementary Information**, Reduced Models). In our model, PL arises from an increase in the precision of representation of the training stimuli, which we instantiated by increasing the gain of the training-driven sensory neurons. When we removed this (**Fig. S5**), the model's behavior was comparable to behavior of observers in the control group (**Fig. S4**). Note, however, that other mechanisms can produce similar effects. In particular, a model variant in which gain modulation was replaced with tuning changes (such that more neurons in the population encode the trained directions) also captured many features of the data (see Supplementary Information; **Fig.**

S6). We focused our analysis on the gain change model, because it fit the data the best, was simpler, and had stronger support from prior empirical literature (16, 24, 52).

When we removed the boundary avoidance component (i.e., warping of the tuning curves), pre- and post-training estimation biases had roughly the right pattern across stimuli and pre- vs. post learning (Fig S7), but were much smaller than those in the human data. On its own, the conditional inference step does produce biased estimates (due to the asymmetric posterior), as found in previous studies (34, 43). Yet, the magnitude of this bias is limited by the likelihood function's width (that arises in our model from a combination of gain, number of neurons, and tuning curve width (74)). We concluded that boundary avoidance was necessary to explain the magnitude of overestimation as well as the misclassified estimates observed far from the boundary.

When we removed the conditional inference (implicit decision) step, the estimate distributions became unimodal — in contrast with bimodal distributions seen in most of the data (Fig. S8). In this model variant, even though the majority of the likelihood function sometimes fell on the side of the discrimination boundary inconsistent with the true stimulus category (as in the full model, Fig. 2A), the mean was much closer to the boundary than if it were conditioned on the implicit categorization judgment. This mean (the estimate) therefore contributed to the left tail of a now unimodal, higher variance estimate distribution, with the tail crossing the boundary (Fig. S8) — instead of contributing to a separate sub-distribution, as when conditional inference is used (Fig. 2A). Thus, conditional inference was necessary to explain the bimodal shape of the estimate distribution.

Quantitative model comparisons confirmed our observations about each of the model variants (see **Supplementary Information**). We compared models using two metrics: one was the goodness-of-fit of each model (i.e., the in-sample loss), and the other was the generalization performance (i.e., the out-of-sample cross-validated loss). Each model variant, including the reduced models and the tuning changes models, provided a worse fit of the data than the full model (Fig. S9A). The generalization performance of the full model, the no-conditional-inference model, and the tuning-change model ("TC") were similarly good, with the other models being much worse (Fig. S9B). However, only the full model and tuning-change model ("TC", which replaced gain changes in the full model with tuning changes) could reproduce the striking bimodal estimation data characteristic of human observers, suggesting that the no-conditional-inference model is inadequate.

Discussion

The signature characteristic of PL is training-induced performance improvement, most commonly associated with a discrimination task (7–11, 29, 32, 33, 44, 53). We asked how training would concurrently affect stimulus appearance. Whereas one might have expected that improvement in discrimination accuracy would be accompanied by reduced estimation biases (i.e., learning should improve veridicality), our data revealed the opposite. Participants who trained on either a discrimination or estimation task showed improvement in discrimination accuracy and coherence thresholds, but also exhibited increases in repulsive biases that were present (and substantial) prior to training (Fig. 1C). Moreover, although training was restricted only the (±4°) directions, these effects were transferred to untrained nearby directions (both ±2° and ±8°), illustrating within-category transfer (8, 10).

The PL effects we observed bear some similarity to the shorter-timescale effects of attention and adaptation. Specifically, spatial attention has been shown to improve behavioral performance by magnifying task-relevant sensory aspects (54, 55), even when that entails creating a 'less veridical' representation of the stimulus (56). At slightly longer timescales, adaptation improves direction discrimination for directions near that of the adaptor while repelling perceived directions away from that of the adaptor (57, 58). Both of these effects are relatively short-lived, and thus do not explain the results of the current study, which are evident across multiple days.

Our experiments measure the effects of PL without feedback, as has been done in a number of previous studies (e.g., (53, 59–61). This is more consistent with the "unsupervised learning" that occurs in many real-world situations. Moreover, a number of authors have also suggested that it is preferable to avoid feedback when using estimation judgements to assess subjective appearance, in the context of perceptual learning (8), category learning (62), interaction of subjective and objective perceptual organizations (63), 3-D form (64), and appearance of various perceptual dimensions (e.g., (65, 66); review by (67)).

Overwhelmingly, PL studies use two-alternative forced choice discrimination tasks and only a handful have used estimation tasks. Most of these (64-66) provided feedback, and reported reduced estimation variability, but did not examine effects on appearance bias. One study used estimation in the absence of feedback, and found that estimation training increased overestimation as well as discrimination accuracy (8). Another PL study, which employed feedback during training, used both estimation and discrimination tasks and analyzed cross-task transfer (68). Training on discrimination had no statistically detectable effect on estimation variance, and training on estimation had no detectable effect on discrimination thresholds. Within-task testing revealed enhanced discrimination performance and reduced estimation variance. This suggests that the absence of feedback was critical for our findings, i.e., robust cross-task transfer, increased overestimation, and reductions in discrimination thresholds.

How did training lead to increased overestimation? Our findings show that PL-induced increases in the precision of sensory encoding can interact with implicit categorization and a non-uniform internal representation to repel percepts away from the discrimination boundary. Interestingly, repulsion in the appearance of near-boundary stimuli resembles a well-known characteristic of category learning – between-category expansion – the repulsive distortions of values near the category boundary (69). Category learning typically enhances discrimination between categories, whereas within-category discrimination is reduced or remains unchanged. For example, in color perception, an object belonging to a group of mostly red objects is judged to be redder than an identically-colored object belonging to another group of mostly violet objects (62). That is, color appearance is distorted toward category means, based on the hue statistics of the two groups. Despite some findings of associations between category learning and PL (9, 10, 68), effects of category learning on appearance (e.g., between-category expansion or within-category compression) had not been explored in PL. Our model demonstrates that distinctions between perceptual categories may be enhanced by changes in sensory encoding (e.g., in gain or tuning) over the course of training, without needing to invoke changes in cognitive or decision-making processes.

Why would observers perform an implicit categorization in the estimation task, when they were not instructed to do so? One possibility is that clockwise and counterclockwise of horizontal motion are "natural categories" (38) that are shaped by the structure of the world. The fact that the tilt of near-

cardinal orientations is perceptually exaggerated also means that observers' percepts of clock-wise and counter-clockwise tilts clump together, forming "natural categories." A real-world example of this is hanging a picture on a wall – a picture titled slightly clockwise off straight will look very askew. These examples provide an intuition for the relation between high sensitivity around "anchor" stimuli, and the automatic grouping of nearby values into categories. Although overestimation has been reported when observers only perform and estimation task (8), another possibility is that observers may be influenced by the experimental design, either because they were asked to perform the discrimination task (albeit in separate blocks of trials) during the pretest, or because the ensemble of presented stimuli were symmetrically distributed around the horizontal boundary.

Alternatively, empirical (70, 71) and theoretical (46, 72) studies suggest that observers may naturally and implicitly commit to a high-level interpretation of a stimulus before estimating it (i.e., conditional inference). This process has been considered as a perceptual analogue of confirmation bias (73). Although such biases may seem detrimental for perception, conditional inference may confer distinct advantages (46), including reducing energy costs (74, 75), optimizing use of neural resources by discarding unnecessary details about a stimulus, and protecting crucial information from internally generated noise by storing it in a discrete format (46). Specifically, such a strategy is advantageous when the observer has a good chance of correctly discriminating a stimulus in the presence of post-decisional noise (43, 46).

Notably, studies of conditional inference have found that increased sensory noise (e.g., controlled by the motion coherence of an RDK) leads to larger estimation biases (43, 45). However, in our study, the stimulus' motion coherence was constant over pre-test, training, and post-test, and noise coherence thresholds decreased with training (a known consequence of perceptual learning, e.g., (29–33)) while estimation biases increased. Thus, our results show increased estimation biases and increased sensitivity after learning, consistent with our model.

Theories of efficient coding may offer a unifying framework relating PL, discriminability, and perceptual biases. The brain devotes more resources to representing features of the environment that are more common. For example, cardinal orientations (horizontal/vertical) are more common than oblique orientations in natural retinal images due to gravity and the orientation of the human body (47). Correspondingly, the brain devotes more sensory neurons to the cardinals and human observers exhibit better discriminability for the cardinals than obliques, but also larger biases away from them (38–41). This is true for many other sensory features, including motion direction, and theories of efficient coding can successfully predict discriminability and estimation biases in human and animal behavior based on environmental statistics (35, 42, 47, 48, 76).

In our model, to account for biases in motion perception that were even present on Day 1, we assumed that the brain devotes more sensory neurons to representing horizontal motion. That is, tuning changes occurred over development due to exposure to non-uniform environmental statistics. Likewise, a model variant that assumes that similar tuning changes additionally occur on the time-scale of days explains the data well. This suggests that a similar mechanism of efficient coding may underlie PL, whether in development or adulthood, which enhances discriminability at the cost of also increasing perceptual biases.

Different neuronal changes may underlie performance improvement in PL. We modeled PL as a change in the gain of sensory neurons representing the trained motion direction. A human fMRI study

demonstrated that PL may strengthen attentional (gain) modulation of sensory representations in cortex (24), an account that has been supported by physiological recordings in cat visual cortex (16) and gerbil auditory cortex (77). Tuning changes in sensory neurons (i.e., sharpening and lateral shifts in individual tuning curves) may also be important for PL (13, 14, 17, 18). We fit a variant of our model in which we replaced gain modulation with tuning changes (see **Supplementary Information**), such that neurons encoding the trained stimuli motion directions would be more densely packed after training (48). This model variant reproduced human behavior nearly as well as our chosen model, consistent with the fact that various neural mechanisms may explain the observed human behavior. Our behavioral data is not sufficient to adjudicate between these possibilities. Our model and data do, however, illustrate that a simple change in sensory neurons may interact with implicit categorization to produce idiosyncratic and unexpected patterns of behavior with learning.

The present findings can have translational implications for real-world manifestations of PL such as perceptual expertise and clinical rehabilitation. For example, when learning to categorize CT images into “cancerous” and “benign” (78–80), a radiologist becomes increasingly sensitive to differences between similar images and better at her job. The discriminating features may become increasingly salient over training, altering the appearance of both cancerous and benign images from the way they initially looked.

In conclusion, we found that PL improves discrimination and biases estimation. To explain these counterintuitive findings, we propose that PL in discrimination tasks may reflect improved categorization, associated with biases in appearance. Our model strengthens the links between category learning and PL (9, 10, 81).

Acknowledgments

This work was supported by a research grant from the NIH RO1 EY016200 to MC, a National Defense Science and Engineering Graduate fellowship to CSB, funding from the National Institute of Psychobiology in Israel to SFAS, and a research grant ISF1198/22 to SFAS. We thank members of the Carrasco Lab, especially Shao-Chin Hung and Marc Himmleberg, as well as Mike Landy for useful comments on early versions of the manuscript.

Author contributions

SS and MC designed research; SS performed empirical research; SS and CB analyzed empirical data; CB and SS designed the computational model and ES and MC supervised the model design; CB implemented the computational model; SS, CB and MC wrote the paper; All authors edited the manuscript.

References

1. D. Sagi, Perceptual learning in Vision Research. *Vision Res.* **51**, 1552–1566 (2011).
2. T. Watanabe, Y. Sasaki, Perceptual learning: Toward a comprehensive theory. *Annu. Rev. Psychol.* **66**, 197–221 (2015).
3. A. R. Seitz, Perceptual learning. *Curr. Biol.* **27**, R631–R636 (2017).
4. H. R. Dinse, P. Ragert, B. Pleger, P. Schwenkreis, M. Tegenthoff, Pharmacological modulation of perceptual learning and associated cortical reorganization. *Science (80-.).* **301**, 91–94 (2003).
5. C.-T. Law, J. I. Gold, Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. *Nat. Neurosci.* **11**, 505–513 (2008).
6. M. W. Chu, W. L. Li, T. Komiyama, Balancing the Robustness and Efficiency of Odor Representations during Learning. *Neuron* **92**, 174–186 (2016).
7. T. Saffell, N. Matthews, Task-specific perceptual learning on speed and direction discrimination. *Vision Res.* **43**, 1365–1374 (2003).
8. S. Szpiro, M. Spering, M. Carrasco, Perceptual learning modifies untrained pursuit eye movements. *J. Vis.* **14**, 1–13 (2014).
9. R. Wang, *et al.*, Perceptual learning at a conceptual level. *J. Neurosci.* **36**, 2238–2246 (2016).
10. Q. Tan, Z. Wang, Y. Sasaki, T. Watanabe, Category-Induced Transfer of Visual Perceptual Learning. *Curr. Biol.* **29**, 1374-1378.e3 (2019).
11. J. Yang, *et al.*, General learning ability in perceptual learning. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 19092–19100 (2020).
12. B. A. Dosher, P. Jeter, J. Liu, Z. L. Lu, An integrated reweighting theory of perceptual learning. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 13678–13683 (2013).
13. A. Schoups, R. Vogels, N. Qian, G. Orban, Practising orientation identification improves orientation coding in V1 neurons. *Nature* **412**, 549–553 (2001).
14. M. Sanayei, *et al.*, Perceptual learning of fine contrast discrimination changes neuronal tuning and population coding in macaque V4. *Nat. Commun.* **9**, 1–15 (2018).
15. T. Van Kerkoerle, S. A. Marik, S. M. Z. A. Borgloh, C. D. Gilbert, Axonal plasticity associated with perceptual learning in adult macaque primary visual cortex. *Proc. Natl. Acad. Sci.* **115**, 10464–10469 (2018).
16. T. Hua, *et al.*, Perceptual Learning Improves Contrast Sensitivity of V1 Neurons in Cats. *Curr. Biol.* **20**, 887–894 (2010).
17. R. . E. . Crist, *et al.*, Learning to see: experience and attention in primary visual cortex. *Nat Neurosci* **4**, 519–25 (2001).
18. W. Li, V. Piëch, C. D. Gilbert, Perceptual learning and top-down influences in primary visual cortex. *Nat. Neurosci.* **7**, 651–657 (2004).
19. W. Li, V. Piëch, C. D. Gilbert, Learning to Link Visual Contours. *Neuron* **57**, 442–451 (2008).
20. K. Shibata, T. Watanabe, Y. Sasaki, M. Kawato, Perceptual Learning Incepted by Decoded fMRI Neurofeedback Without Stimulus Presentation. *Science (80-.).* **334** (2011).
21. J. A. Diaz, F. Queirazza, M. G. Philiastides, Perceptual learning alters post-sensory processing in human decision-making. *Nat. Hum. Behav.* **1**, 1–9 (2017).
22. K. Jia, *et al.*, Recurrent Processing Drives Perceptual Plasticity. *Curr. Biol.* **30**, 4177-4187.e4 (2020).
23. Y. Yotsumoto, *et al.*, Different Dynamics of Performance and Brain Activation in the Time Course of

Perceptual Learning. *Neuron* **57**, 827–833 (2008).

24. A. Byers, J. T. Serences, Enhanced attentional gain as a mechanism for generalized perceptual learning in human visual cortex. *https://doi.org/10.1152/jn.00353.2014* **112**, 1217–1227 (2014).

25. X. Wang, Y. Zhou, Z. Liu, Transfer in motion perceptual learning depends on the difficulty of the training task. *J. Vis.* **13**, 5- (2013).

26. B. Dosher, Z. L. Lu, Visual Perceptual Learning and Models. *Annu. Rev. Vis. Sci.* **3**, 343–363 (2017).

27. V. R. Bejjanki, J. M. Beck, Z.-L. L. Lu, A. Pouget, Perceptual learning as improved probabilistic inference in early sensory areas. *Nat. Neurosci.* **14**, 642–648 (2011).

28. G. Sotiropoulos, A. R. Seitz, P. Seriès, Performance-monitoring integrated reweighting model of perceptual learning. *Vision Res.* **152**, 17–39 (2018).

29. B. A. Dosher, Z. L. Lu, Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 13988–13993 (1998).

30. R. W. Li, D. M. Levi, S. a Klein, Perceptual learning improves efficiency by re-tuning the decision “template” for position discrimination. *Nat. Neurosci.* **7**, 178–183 (2004).

31. L.-Q. Q. Xiao, *et al.*, Complete Transfer of Perceptual Learning across Retinal Locations Enabled by Double Training. *Curr. Biol.* **18**, 1922–1926 (2008).

32. I. Donovan, S. F. A. Szpiro, M. Carrasco, Exogenous attention facilitates location transfer of perceptual learning. *J. Vis.* **15**, 1–16 (2015).

33. S. C. Hung, M. Carrasco, Feature-based attention enables robust, long-lasting location transfer in human perceptual learning. *Sci. Rep.* **11**, 17293 (2021).

34. A. A. Stocker, E. P. Simoncelli, A Bayesian Model of Conditioned Perception. *Adv. Neural Inf. Process. Syst.* **20**, 1409–1416 (2008).

35. X. X. Wei, A. A. Stocker, Lawful relation between perceptual bias and discriminability. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 10244–10249 (2017).

36. Y. Kamitani, F. Tong, Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.* **8**, 679–685 (2005).

37. B. ‘Brit’ Brogaard, *et al.*, The real epistemic significance of perceptual learning. *Inquiry* **61**, 543–558 (2018).

38. S. Appelle, Perception and discrimination as a function of stimulus orientation: the “oblique effect” in man and animals. *Psychol. Bull.* **78**, 266–278 (1972).

39. H. J. Rauber, S. Treue, Reference repulsion when judging the direction of visual motion. *Perception* **27**, 393–402 (1998).

40. G. Loffler, H. S. Orbach, Anisotropy in judging the absolute direction of motion. *Vision Res.* **41**, 3677–3692 (2001).

41. A. E. Krukowski, *et al.*, Human discrimination of visual direction of motion with and without smooth pursuit eye movements. *J. Vis.* **3**, 831–840 (2003).

42. X. Xu, C. E. Collins, I. Khaytin, J. H. Kaas, V. A. Casagrande, Unequal representation of cardinal vs. oblique orientations in the middle temporal visual area. *Proc. Natl. Acad. Sci.* **103**, 17490–17495 (2006).

43. L. Luu, A. A. Stocker, Post-decision biases reveal a self-consistency principle in perceptual inference. *Elife* **7** (2018).

44. A. R. Seitz, T. Watanabe, Is subliminal learning really passive? *Nature* **422**, 2003 (2003).

45. M. Jazayeri, J. A. Movshon, A new perceptual illusion reveals mechanisms of sensory decoding. *Nature* **446**, 912–915 (2007).

46. C. Qiu, L. Luu, A. A. Stocker, Benefits of Commitment in Hierarchical Inference. *Psychol. Rev.* (2020)

https:/doi.org/10.1037/REV0000193 (July 19, 2021).

47. A. R. Girshick, M. S. Landy, E. P. Simoncelli, Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nat. Neurosci.* **14**, 926–932 (2011).

48. D. Ganguli, E. P. Simoncelli, Efficient Sensory Encoding and Bayesian Inference with Heterogeneous Neural Populations. *Neural Comput.* **26**, 2103–2134 (2014).

49. W. J. Ma, J. M. Beck, P. E. Latham, A. Pouget, Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9**, 1432–1438 (2006).

50. D. Ganguli, E. P. Simoncelli, Implicit encoding of prior probabilities in optimal neural populations. *Adv. Neural Inf. Process. Syst.* **2010**, 658–666 (2010).

51. P. Seriès, A. A. Stocker, E. P. Simoncelli, Is the Homunculus “Aware” of Sensory Adaptation? *Neural Comput.* **21**, 3271–3304 (2009).

52. M. L. Caras, D. H. Sanes, Top-down modulation of sensory cortex gates perceptual learning. *Proc. Natl. Acad. Sci.* **114**, 9972–9977 (2017).

53. K. Ball, R. Sekuler, Direction-specific improvement in motion discrimination. *Vision Res.* **27**, 953–965 (1987).

54. S. Ling, T. Liu, M. Carrasco, How spatial and feature-based attention affect the gain and tuning of population responses. *Vision Res.* **49**, 1194–1204 (2009).

55. A. Fernández, S. Okun, M. Carrasco, Differential Effects of Endogenous and Exogenous Attention on Sensory Tuning. *J. Neurosci.* **42**, 1316–1327 (2022).

56. V. Mehrpour, J. C. Martinez-Trujillo, S. Treue, Attention amplifies neural representations of changes in sensory input at the expense of perceptual accuracy. *Nat. Commun.* **11**, 2128 (2020).

57. P. R. Schrater, E. P. Simoncelli, Local velocity representation: evidence from motion adaptation. *Vision Res.* **38**, 3899–3912 (1998).

58. C. W. G. Clifford, Perceptual adaptation: motion parallels orientation. *Trends Cogn. Sci.* **6**, 136–143 (2002).

59. S. Koyama, A. Harner, T. Watanabe, Task-dependent changes of the psychophysical motion-tuning functions in the course of perceptual learning. *Perception* **33**, 1139–1147 (2004).

60. M. H. Herzog, M. Fahle, The role of feedback in learning a vernier discrimination task. *Vision Res.* **37**, 2133–2141 (1997).

61. A. A. Petrov, B. A. Dosher, Z. L. Lu, Perceptual learning without feedback in non-stationary contexts: Data and model. *Vision Res.* **46**, 3177–3197 (2006).

62. R. L. Goldstone, Effects of Categorization on Color Perception. *Psychol. Sci.* **6** (1995).

63. M. Carrasco, I. Chang, The interaction of objective and subjective organizations in a localization search task. *Percept. Psychophys. 1995 578* **57**, 1134–1150 (1995).

64. M. L. Braunstein, J. T. Todd, On the Distinction Between Artifacts and Information. *J. Exp. Psychol. Hum. Percept. Perform.* **16**, 211–216 (1990).

65. M. Carrasco, S. Ling, S. Read, Attention alters appearance. *Nat. Neurosci.* **7**, 308–313 (2004).

66. K. Anton-Erxleben, K. Herrmann, M. Carrasco, Independent effects of adaptation and attention on perceived speed. *Psychol. Sci.* **24**, 150–9 (2013).

67. M. Carrasco, A. Barbot, Spatial attention alters visual appearance. *Curr. Opin. Psychol.* **29** (2019).

68. S. C. Green, F. Kattner, M. H. Siegel, D. Kersten, P. R. Schrater, Differences in perceptual learning transfer as a function of training task. *J. Vis.* **15**, 1–14 (2015).

69. R. L. Goldstone, Y. Lippa, R. M. Shiffrin, Altering object representations through category learning. *Cognition* **78**, 27–43 (2001).

70. S. Ding, C. J. Cueva, M. Tsodyks, N. Qian, Visual perception as retrospective Bayesian decoding from high- to low-level features. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E9115–E9124 (2017).

71. E. Zamboni, T. Ledgeway, P. V. McGraw, D. Schluppeck, Do perceptual biases emerge early or late in visual processing? Decision-biases in motion perception. *Proc. R. Soc. B Biol. Sci.* **283**, 20160263 (2016).

72. S. Wu, H. Lu, A. Lee, A. Yuille, Motion Integration Using Competitive Priors. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **5604 LNCS**, 235–258 (2009).

73. R. S. Nickerson, Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **2**, 175–220 (1998).

74. S. J. Gershman, E. J. Horvitz, J. B. Tenenbaum, Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science (80-.).* **349**, 273–278 (2015).

75. H. A. Simon, On the behavioral and rational foundations of economic dynamics. *J. Econ. Behav. Organ.* **5**, 35–55 (1984).

76. X.-X. Wei, A. A. Stocker, A Bayesian observer model constrained by efficient coding can explain “anti-Bayesian” percepts. *Nat. Neurosci.* **18**, 1509–1517 (2015).

77. M. L. Caras, D. H. Sanes, Top-down modulation of sensory cortex gates perceptual learning. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9972–9977 (2017).

78. G. Dale, A. Cochrane, C. S. Green, Individual difference predictors of learning and generalization in perceptual learning. *Attention, Perception, Psychophys.*, 1–15 (2021).

79. S. M. Frank, *et al.*, Supervised Learning Occurs in Visual Perceptual Learning of Complex Natural Images. *Curr. Biol.* **30**, 2995-3000.e3 (2020).

80. S. Waite, *et al.*, Analysis of perceptual expertise in radiology – Current knowledge and a new perspective. *Front. Hum. Neurosci.* **13**, 213 (2019).

81. F. Kattner, C. R. Cox, C. S. Green, Transfer in rule-based category learning depends on the training task. *PLoS One* **11** (2016).

82. M. S. Landy, M. S. Banks, D. C. Knill, “Ideal-Observer Models of Cue Integration” in *Sensory Cue Integration*, (Oxford University Press, 2011), pp. 5–29.

83. S. Shen, W. J. Ma, Variable precision in visual perception. *Psychol. Rev.* **126**, 89–132 (2019).

84. D. Rahnev, R. N. Denison, Suboptimality in perceptual decision making. *Behav. Brain Sci.* **41**, e223 (2018).

85. A. H. Yoo, L. Acerbi, W. J. Ma, Uncertainty is maintained and used in working memory. *J. Vis.* **21**, 13 (2021).

86. L. T. Maloney, H. Zhang, Decision-theoretic models of visual perception and action. *Vision Res.* **50**, 2362–2374 (2010).

87. A. E. Orhan, C. R. Sims, R. A. Jacobs, D. C. Knill, The Adaptive Nature of Visual Working Memory: *https://doi.org/10.1177/0963721414529144* **23**, 164–170 (2014).

88. L. Acerbi, W. J. Ma, S. Vijayakumar, A Framework for Testing Identifiability of Bayesian Models of Perception in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Q. Weinberger, Eds. (Curran Associates, Inc., 2014).

89. T. D. Albright, Direction and orientation selectivity of neurons in visual area MT of the macaque. *https://doi.org/10.1152/jn.1984.52.6.1106* **52**, 1106–1130 (1984).

90. M. Jazayeri, J. A. Movshon, Optimal representation of sensory information by neural populations. *Nat. Neurosci.* **9**, 690–696 (2006).

91. D. Sejdinovic, B. Sriperumbudur, A. Gretton, K. Fukumizu, Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Stat.* **41**, 2263–2291 (2013).

92. Z. Szabó, Information theoretical estimators toolbox. *J. Mach. Learn. Res.* **15.1**, 283–287 (2014).

93. R. T. Marler, J. S. Arora, Survey of multi-objective optimization methods for engineering. *Struct. Multidiscip. Optim.* **26**, 369–395 (2004).

94. L. Acerbi, W. J. Ma, Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search in *Advances in Neural Information Processing Systems*, I. Guyon, *et al.*, Eds. (Curran Associates, Inc., 2017).

Methods

Observers

Twenty-three human adults participated (mean age = 28.9, SD = 1.9; nine males), after giving their written informed consent. All had normal or corrected to normal vision, were untrained and did not know the experiment's purpose. Two participants had to be omitted from the training: one could not perform the estimation task during pretest (average estimation deviated more than 70º from veridical); the other participant's performance on discrimination did not differ from chance, leading to a total of twenty-one observers.

Visual stimuli and display

Stimuli were random dot kinematograms (RDK) with dots moving at 15º/s in a stationary aperture with a 5º radius, sparing 0.75º around the central fixation cross. On each frame each dot was assigned a direction that was either the coherent direction or a different random direction with the probability matching the coherence level found for each observer. Dots were black (4 pixels, 3cd/m2) and were shown on a uniform gray background, with dot density 1.65 dots per square degree. Stimuli were displayed on a calibrated 41x30 cm CRT monitor (IBM P260) with a resolution of 1280x960 pixels, and a 100 Hz refresh rate. Observers were seated at 57 cm distance from the screen with their head supported by a combined chin- and forehead-rest.

Testing and training tasks

The experiment consisted of five 60-min sessions, one per day, over five consecutive days; the first (pretest) and last (posttest) sessions were identical, and the three intermediate sessions were training sessions (**Fig. S1**). The pretest included a short practice on both the estimation and the discrimination tasks. For each observer we tested discrimination coherence thresholds for motion directions of ±4º using three randomly interleaved 60-trial 3-down-1-up staircases. We estimated coherence thresholds by averaging the thresholds reached by the three staircases. This coherence level was then used for all following testing and training.

During pre- and posttest sessions, we measured performance on six randomly presented coherent motion directions (directions relative to horizontal to the right): -8º, -4º, -2º (downwards from horizontal) and 8º, 4º, 2º (upwards from horizontal). Testing sessions (pretest and posttest) included a block of motion discrimination and a block of direction estimation; the order of the two blocks was counterbalanced across observers. Each block consisted of 360 trials. In training sessions, observers were trained either on the estimation task or in the discrimination task, only directions of ±4º directions were presented, and each session consisted of 720 trials presented in four blocks.

Each trial started with a 500-ms fixation cross at the center of the screen, then the 200-ms RDK appeared followed by a 700-ms ISI after which an auditory start signal indicated that a response could be given (**Fig. 1A**). In discrimination blocks, observers pressed a key on the keyboard indicating upward or downward motion relative to horizontal, and the response had to be given within 900 ms. In estimation blocks, a randomly oriented line appeared and, using a mouse, observers were given 4-seconds to orient the line according to the motion direction they perceived; the initial orientation of the line varied uniformly around horizontal with a variance of 5º. Observers mostly responded within the timeline (98% of trials). No feedback was given either for discrimination or estimation tasks as feedback is not necessary for PL to occur and could affect reports rather than appearance (see Discussion).

Statistical Analysis

We used a repeated measures analysis of variance (ANOVA), using session (pretest/posttest) and directions (±2º, ±4º or ±8º) as within-subject factors, and group or training conditions as between-subject factors. When assumptions for sphericity were not met, results were corrected using Greenhouse-Geisser. For both estimation and discrimination tasks, up and down responses were combined: we negated directions and responses for the downward trials and merged the data with that of the upward trials.

Modeling Methods

We hypothesized that PL in our task can be explained by increases in the precision of the internal sensory representation — a representation that is already warped to devote more resources to certain features. To test this hypothesis, we created a probabilistic observer model that formalizes each of its commitments. Probabilistic observer models (e.g., Bayesian observer models) are most often used to describe how an observer should behave in a task to optimize their performance, given some variability (e.g., uncertainty, ambiguity, sensory noise) that interferes with decision-making (82, 83). Such models rely on the assumption that the observer (or some part of their brain) is “aware” of these noise sources, e.g., the distribution of a stimulus’s value across trials, the brain’s noisy measurement of the stimulus, and so on (51). However, it is straightforward to modify the model such that the observer has some set of incomplete or incorrect beliefs (84, 85). These “imperfectly optimal observers” (86) can reproduce idiosyncrasies in human behavior while maintaining the fundamental commitment that observers account for their own sensory uncertainty in decision-making rather than ignoring it (49, 87, 88). Such observer models can account for complexities in human behavior, without being beholden to explaining how each step is part of a normative account or optimal solution. We propose one such observer model, consisting of encoding and decoding stages, which “learns” and performs both the discrimination and estimation tasks.

Task Structure

On every trial, there is a 0.5 probability of the stimulus motion direction being CW of horizontal, expressed as $p(C = CW) = 0.5$, where C, the category of the stimulus, takes the value CCW or CW. The overall stimulus distribution, $p(s)$, is the normalized sum of six delta functions at ±2, ±4, and ±8º, i.e., the probability of each of the six possible motion directions is equal. Once the category is drawn on each trial, the new, “category-conditioned” stimulus distribution $p(s|C)$ is the half of $p(s)$ on the side of the discrimination boundary (0º) that is consistent with the category C.

Encoding

We assume the stimulus is measured ("encoded") by a population of N = 10 neurons whose tuning curves tile the space of motion directions but are warped such that more neurons represent motion directions near the horizontal, consistent with physiological measurements and theoretical studies of efficient encoding (48, 50). The tuning curves specify each neuron's mean firing rate as a function of stimulus direction. Each tuning curve, before warping, is a single cycle of a cos^2 function raised to a power (0.3745) such that its full-width at half maximum (FWHM), w_t, is 71º (48), consistent with typical tuning of motion-direction-selective neurons in macaque area MT (89). Each neuron has a baseline firing rate, b, and a gain, g, which specifies the above-baseline firing rate for the preferred stimulus (i.e., the maximum response). The pre-training gains, g_{pre}, are assumed to be identical across the population. The post-training gains are elevated according to a Gaussian profile across the population, centered at 0º, with a standard deviation equal to the spacing between neurons (specifically, with ten neurons, the standard deviation is 4º). A parameter g_{post} specified the gain of the most responsive neuron in the population. This captures our assumption that learning modulates the gain of sensory neurons, relative to their sensitivity to the trained motion directions.

Warping in our model embodies the efficient coding notion of devoting more neurons and spikes to more common stimuli (48, 50). This leads to a behavior that resembles avoidance of the boundary when estimating the stimulus, so we refer to this as the "boundary avoidance" component of the model. The neural population is formed from a set homogeneous ("convolutional") set of tuning curves that are re-mapped by warping the motion direction axis according to a parametric function that is fit to the data of each observer, but assumed unchanged throughout the experiment (i.e., unaffected by learning). The warping function is defined as the cumulative integral of a cell "density" function: $h(s,a,w_b,f) = a + 2 + \phi(s,-w_b/2,f) + \phi(-s,w_b/2,f)$, where ϕ is the cumulative Gaussian function. Motion direction is designated with variable s, and the shape of the function is governed by three parameters: an amplitude, a, that controls the overall magnitude of the warping (i.e., how "extremely" the observer avoids the boundary); w_b, the width of the region over which tuning curves are drawn toward the boundary; and f, the standard deviation of the cumulative Gaussians, which controls the fuzziness of the template's edges, leading to warping that is more graded or discrete. Examples are shown in **Fig. S10.** Before usage, the function h is numerically normalized to integrate to 1.

For a stimulus direction s, each neuron's spike count r_i, is drawn independently from a Poisson distribution, with rate determined by its tuning curve evaluated at the (warped) value of s. The noisy "population response" (i.e., the vector of spike counts from each neuron), denoted **r**, is a sample of the "measurement distribution" $p(\mathbf{r}|s)$.

Decoding

The internal representation of the stimulus **r** must be "read out" or decoded to discriminate or estimate motion directions. For this purpose, we used a *maximum a posteriori* or MAP decision rule for the discrimination judgment, and Bayes-least-squares rule for the estimation task. Our methods are identical to those reported in other studies (49, 90) so we refer the reader to their methods for details, and simply provide an intuition here.

Both decoders rely on a likelihood function, in which the measurement distribution is expressed as a function of the stimulus, s, for each noisy population response $\mathbf{r}$. For our model, the log likelihood is equal to the sum (across neurons) of the log of each tuning curve, each weighted by the observed

spike count of the associated neuron (49, 90). We incorporate one additional assumption, that the decoder is 'unaware' of the warping in the encoder, and so assumes a homogeneous ("convolutional") encoding population in computing the likelihoods. This idea has been used successfully in the domain of adaptation to explain biases in perception (51).

The likelihood fluctuates randomly on each trial, due to the variability of **r**. In the case when w_b or a is zero, the likelihood is symmetric and centered on the true stimulus motion direction, i.e., its mean is an unbiased estimator of the true stimulus (equivalent to probabilistic population coding; (49, 90). Its width, which corresponds with the observer's uncertainty about the motion direction generating the internal measurement **r**, depends on gain and baseline firing rate. If w_b and a are larger than zero, the likelihood is stretched away symmetrically in half around the boundary a distance depending on w_b, and suppressed near the boundary an amount depending on a. The likelihood is therefore asymmetric, and the shape of its tails depends on f.

In the discrimination task, the decision variable d represents the log posterior ratio, and is determined by the proportion of the likelihood function that falls on each side of the discrimination boundary: If most of the likelihood mass falls on the positive side of the boundary, $d > 0$, the observer reports "CW" (and vice versa). The distribution of discrimination responses across trials is $p(C_{est}|C)$ and the discrimination accuracy $p(Correct) = p(d>0|C=CW)/2 + p(d<0|C=CCW)/2$. C_{est} denotes the estimated category. Note that d inherits its variability from the population response **r**. To model lapses in the discrimination judgement, we multiply $p(Correct)$ by the same factor $1-\lambda$ for all stimulus motion directions. Note that, in our model fitting, λ is never allowed to go below 0.5.

In the estimation task, the estimate is computed as the expected value of the posterior distribution, i.e., the Bayes-least-squares estimate (34). The posterior is computed on each trial by imposing an implicit discrimination judgment using the rule described above, using the response (CW or CCW) as a conditional prior, $p(s|C_{est})$, which has a value of 1 on the side of the discrimination boundary corresponding to the chosen category, and 0 on the other side. The product of this conditional prior $p(s|C_{est})$, and the likelihood $p(\mathbf{r}|s)$, is the posterior $p(s|\mathbf{r},C_{est})$, and the estimation response on a single trial is the mean of this posterior. The distribution of estimates across trials $p(s_{est}|s)$ is computed numerically for each stimulus, ±2, 4, and 8º, via Monte Carlo simulation. To model motor noise in the execution of the response (orienting an arrow with a mouse), i.i.d. Gaussian noise with standard deviation σ_m is added to each estimate distribution.

Parameter Estimation

The model has six parameters that are fit to individual observers: a, w_b, f, g_{pre}, g_{post}, λ, and five parameters that are shared across observers b, w_t, N, σ_m, and σ_g. We estimated the individual parameters by minimizing a loss function L(**Θ**) (with **Θ** a vector containing the parameters) expressing the fit between each observer's data and model behavior. L(**Θ**) was defined as the weighted sum of two terms: (1) absolute difference (L1 norm) between the discrimination accuracy in the data vs. model behavior (i.e., sum of L1 differences across stimuli and pre- vs. post-learning); and (2) the energy distance (91) between the estimate distributions for the data vs. model (sum of distances across stimuli and pre- vs. post-learning). This distance metric takes into account the entire shape of the estimate distribution and is appropriate for comparing distributions with complex shapes (i.e., like the bimodal distributions we observed). We computed the energy distance as implemented in the Information Theoretic Estimators Toolbox in Matlab (92). The weights on the two terms were used to

rescale the two terms in the loss function into a similar numerical range. These weights were computed by evaluating the objective function on an initial random set of parameters, as is common practice in multi-objective optimization (93). The loss function was stochastic, varying between each run of simulated trials due to sampling variability in the Poisson spike generation. We set the number of simulated trials to 1500 and repeated the optimization 10 times with a new set of initial random parameters.

We fit all free parameters simultaneously for each observer separately using an optimization algorithm designed for stochastic objective functions (94). We ran it 10 times per observer, chose the iteration with the lowest loss. The resulting best-fit parameters were then used to evaluate the model using 1500 simulated trials, generating (stochastic) model behavior for the estimation and discrimination tasks (**Fig. 2B,C**, **Fig. 3**).

Note that, for each observer, we fit the 6 free parameters with approximately 486 data points (80 estimation responses x 2 pre/post-training x 3 motion directions, plus 1 discrimination accuracy x 2 pre/post-training x 3 motion directions), so this is a well-constrained optimization problem (as compared to fitting only the estimation biases for the 6 stimulus directions).

Model Comparison

We compared models using iterated *k*-fold (*k* = 3) cross-validated loss. For each observer and each model, we averaged the CV-loss across folds and iterations (folds x iterations = 102), subtracted this from the average CV-loss from the full model, and then averaged this across observers to quantify each model's generalizability relative to the full model (**Fig. S9**). We also did the same with the average in-sample (training set) loss across observers to quantify goodness-of-fit relative to the full model. To perform the cross-validated model fitting, for each observer, we computed one normalization factor for each objective (estimation and discrimination) based on a fixed initial set of parameters and kept these normalization factors fixed across models. Note that these factors were quite similar across observers, but not identical. This was essential to keep the training and test losses in the same space across models and hence comparable.