# Gaze reflects confidence in explicit decisions while relying on a distinct computational mechanism

Yonatan Stern[1], Ophir Netzer[1], Danny Koren[1], Yair Zvilichovsky[1], Uri Hertz[1,2,3], Roy Salomon[1,2,3*]

[1] School of Psychological Sciences, University of Haifa.

[2] The Integrated Brain and Behavior Research Center (IBBRC), University of Haifa.

[3] Department of Cognitive Science, University of Haifa.

* Corresponding Author: Roy Salomon, royesal@gmail.com

**Keywords:** decision making, metacognition, reinforcement learning, oculomotor behavior, virtual reality

**This PDF file includes:**

Main Text
Figures 1 to 5
Tables 1

**Abstract**

The eyes are considered a 'window' into the mind, shedding light on the cognitive processes leading to explicit decisions. Yet, eye movements are also a distinct oculomotor decision that is the outcome of an independent cognitive system. When exploring novel environments, the convergence and divergence of gaze and explicit decisions highlights the eyes' dual role. However, the computational mechanisms underlying this interplay and its contribution to adaptative behavior remains unclear. Combining virtual reality based associative learning paradigms with computational modeling across one exploratory and two pre-registered experiments (Total N = 115), we examine learning as multi-system process encompassing explicit decisions and ocular expectations. The two exhibited a partial overlap in their learning. On the one hand, participants' explicit predictions about target location and their subsequent anticipatory gaze direction robustly converged, and gaze reveals the latent probability of the explicit choice. Moreover, gaze exhibited computational hallmarks of confidence, reflecting an oculomotor assessment of the explicit prediction. On the other hand, ocular and explicit decisions consistently diverged indicating an independent computation reflected in gaze. Although the two systems learn and ascribe value similarly, they utilize different decision mechanisms. Oculomotor decisions were less driven by choice consistency across trials enabling increased exploration. Under uncertainty explicit and oculomotor decisions tended to diverge, enabling a hedging of one's decision. These findings highlight how together, explicit and oculomotor systems' learning enables adaptive embodied behavior.

**Significance Statement**

The eyes often reveal additional information beyond one's explicit decision, offering a 'window into the mind' (e.g., leading poker players to wear sunglasses). Yet, eye movements are the product of a distinct system, occasionally diverging from explicit decisions. However, what additional information is conveyed by gaze, and what leads to its occasional divergence remain unclear. Using virtual reality and computational modeling, this study demonstrates a partial overlap between the two. Generally, they robustly converge, and gaze reflects a confidence-like assessment of explicit decisions. Yet they also diverge due to distinct decision mechanisms. Their divergence enables a 'hedging' of one's decision under uncertainty. This demonstrates a multimodal decision process, where different action systems make overlapping but independent choices.

## Introduction

Seminal work highlights how distinct cognitive systems interact and shape human learning and behavior in unexpected manners (Kahneman, 2017). We learn through experience, updating both our explicit predictions and internal models of the environment that guide our implicit sensorimotor behavior (Bar, 2004; Helmholtz, 1924). The learned expectations of the two systems typically converge seamlessly. For instance, in the everyday task of checking for oncoming traffic before crossing the street, your oculomotor behavior (e.g., looking to the left for incoming cars) likely converges with your explicit prediction of the oncoming traffic's direction, if asked. Yet in certain instances the two may diverge. Visiting England, despite your ability to update your explicit predictions and know the direction of oncoming traffic correctly, you may nonetheless look in the wrong direction when crossing. The interplay of explicit decisions and implicit sensorimotor behavior is ubiquitous, affecting diverse domains such as perception (Chalk et al., 2010; Girshick et al., 2011), memory (Gilboa & Marlatte, 2017; Squire, 2004; Turk-Browne et al., 2006) and motor control (Applebaum et al., 2024; Maresch et al., 2021; Nielsen, 1963; Salomon et al., 2021; Synofzik et al., 2008). However, little is known regarding the convergence and divergence of explicit predictions and implicit oculomotor expectations during learning.

In complex environments like crossing a busy street, we primarily focus our gaze and attention on the elements we have learned to be important (Leong et al., 2017; Niv, 2019; Niv et al., 2015). Gaze and attention are closely coupled (Deubel & Schneider, 1996; Rizzolatti et al., 1994) and share common neural substrates (Corbetta et al., 1998; Krauzlis et al., 2013; Kustov & Lee Robinson, 1996). Seminal work (Corbetta & Shulman, 2002; Posner et al., 1980) highlights how attention and gaze are directed by exogenous (i.e. stimulus driven) and endogenous (i.e. goal driven) factors such as an arrow within an experimental paradigm (Posner, 1980), statistical regularities inherent in naturalistic environments, and explicitly learned complex knowledge (for reviews see e.g. Nobre & Kastner, 2014; Wolfe, 2020). Ocular expectations are implicitly embedded within gaze dynamics. For example, viewing natural scenes we experience them

effortlessly, however unbeknown to us, distinctive gaze patterns reveal implicit expectations of scene composition (Biederman et al., 1973; Torralba et al., 2006), enabling efficient scanning and processing of information. Furthermore, gaze dynamics shed light on latent cognitive processes such as choice preferences prior to an explicit decision (Krajbich et al., 2010; Krajbich & Rangel, 2011), the reinstatement of objects in memory (Altmann, 2004; Johansson & Johansson, 2014; Van Ede, Chekroud, & Nobre, 2019) and the representation of abstract concepts (Viganò et al., 2024). Thus, gaze dynamics reveal our implicit models of the environment which are typically closely coupled to our explicit decisions and knowledge. Despite significant progress in the field of associative learning (Leong et al., 2017; Vossel et al., 2014), the mechanisms behind the formation of ocular expectations and their connection to learning and the updating of explicit predictions remain unclear.

Updating one's expectations following unexpected events in a volatile environment often affects one's 'feeling of knowing'. When you venture into the street upon your arrival in England, even if you look in the "right" direction you may feel uncertain and hesitant about crossing the street compared to crossing the street outside your home. Confidence, defined as the ability to assess the accuracy of one's decisions, adaptively modulates behavior in a changing environment (Fleming, 2024; Nelson, 1984). Recent studies examining confidence during probabilistic learning found that it tracks the precision of the likelihood distribution (Meyniel et al., 2015), reflects the crossing of an internal boundary (Hertz et al., 2018), and modulates subsequent information seeking behavior (Desender et al., 2018) and reward maximization strategies (Van den Berg et al., 2016). A confidence weighted learning mechanism in which high confidence errors cause a greater modulation of the learning rate than low confidence errors has been proposed (Drugowitsch et al., 2019). Thus, confidence entails additional information pertaining to one's decision. On the one hand it converges robustly with decision accuracy, differentiating correct and incorrect trials (i.e., positive metacognitive sensitivity; Fleming & Lau, 2014; Maniscalco & Lau, 2012; Metcalfe & Shimamura, 1994; Rahnev et al., 2020). On the other hand, it diverges in important manners, reflecting parallel or additional post-decision computational processes (Balsdon et al.,

2020; Fleming, 2024; Fleming & Daw, 2017), integrating prior information in a more optimal manner (Constant et al., 2023) and exhibiting unique biases (Shekhar & Rahnev, 2021). Thus, confidence highlights how different aspects of one's decision reflect the interplay between different cognitive learning systems.

Within the field of confidence both explicit and implicit measures are used. Although explicit confidence ratings are typically used, confidence's relation to implicit measures such as reaction time has long been recognized (Rahnev et al., 2020; Weidemann & Kahana, 2016). Recent work has demonstrated that sensorimotor aspects of one's behavior, such as speech prosody (Goupil et al., 2021; Goupil & Aucouturier, 2021) and motor kinematics (Dotan et al., 2018; Faivre et al., 2020) are related to explicit confidence. Within the ocular modality, extensive research has linked changes in pupil size to internal mechanisms underlying decision certainty (Nassar et al., 2012). Pre-decision pupil dilation is increased for trials that were subsequently rated with low confidence (Ting et al., 2023), and post-decision pupil dilation exhibits the computational signatures of confidence (Colizoli et al., 2018; Urai et al., 2017), presumably reflecting internal evidence accumulation processes (Balsdon et al., 2020). In contrast to pupillometry, the relation between gaze patterns and explicit confidence ratings in humans has not been as extensively studied.

In the current study we characterize the relation between explicit predictions and implicit ocular expectations during associative learning. Across three experiments, participants performed an associative learning task in which they explicitly predicted the location of the upcoming target based on a cue and the learning of the underlying probabilistic rule (see Fig. 1A). Participants were informed that the cue predicts the target's location in 75% of the trials, and that the underlying rule associating the cue and target can be reversed unbeknown to them throughout the experiment. Leveraging eye tracking allowing the measurement of gaze direction prior to the target's appearance, we derive post-decision *ocular expectations* regarding the target's location that can be compared to their *explicit prediction* obtained via a button press of the target's location. Focusing on their convergence and divergence of the two learning systems, we can

broadly outline a hypothesis space that consists of three possibilities (see Fig. 1B& C): (1) *explicit predictions* and *ocular expectations* are independent of each other and fully diverge, (2) The two systems are equivalent, and their learned expectations completely converge, (3) there is a partial overlap between the two learning systems exhibiting both robust convergence and substantial divergence.

In the following analyses, we first examine the convergence of the two systems. As expected, ocular expectations converged robustly with explicit predictions and latent learning processes. Furthermore, the degree of convergence between gaze and explicit predictions exhibited the statistical hallmarks of confidence in one's explicit response. Directly examining this relation in Experiment 3 which included explicit confidence ratings, we found a significant but modest overlap between the two. Turning to examine their divergence, we compare explicit and ocular learning characteristics. Explicit responses were significantly more accurate and with greater metacognitive sensitivity. Delving into their underlying computational learning mechanisms, the two share a similar learning model yet differ in their decision model. Explicit responses exhibit increased response perseverance across trials and ocular responses are more exploratory. Finally, examining the possible synergistic benefits of two partially overlapping learning systems, we found that the systems' responses tend to differ under higher uncertainty, enabling the "hedging" of one's response under uncertainty.

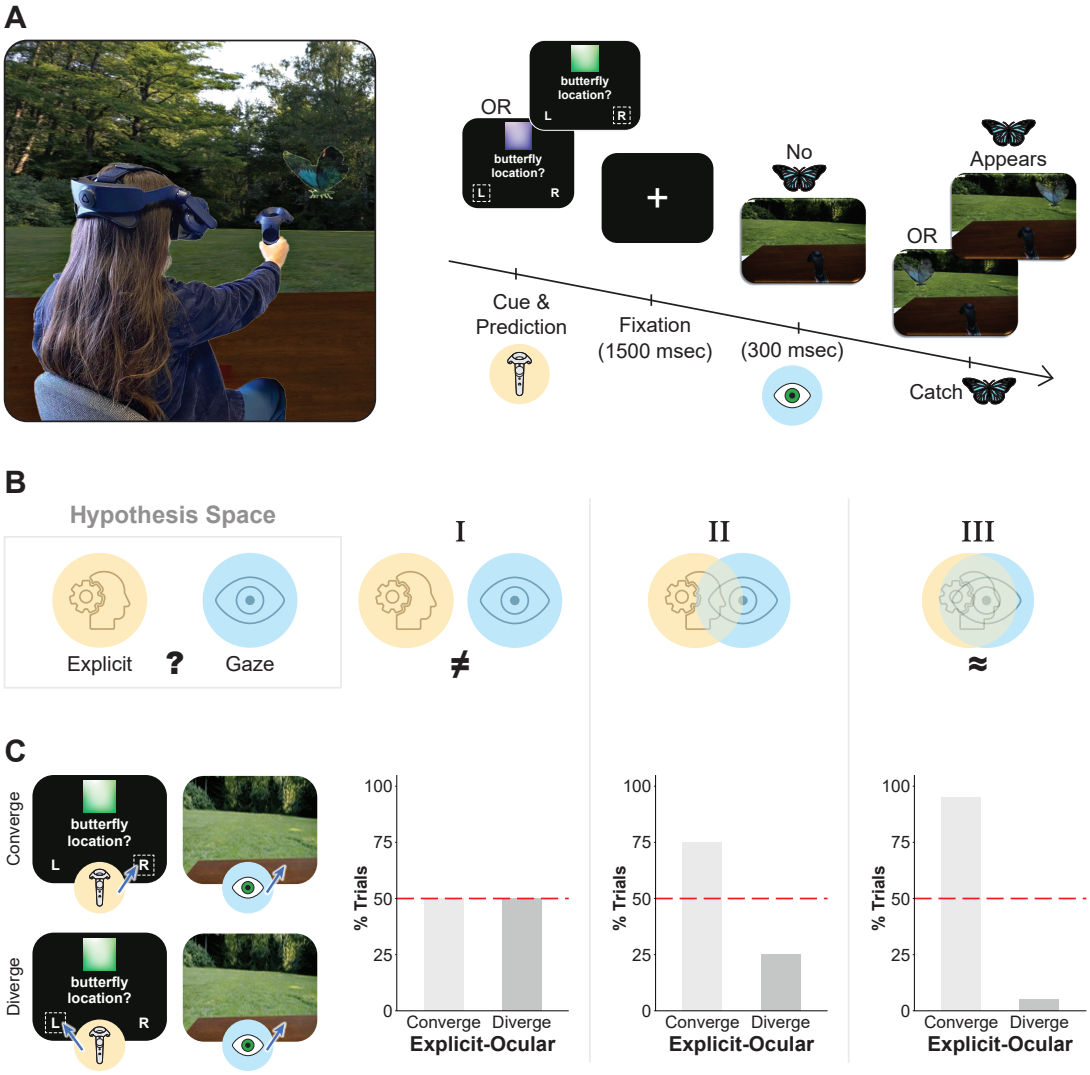


**Figure 1. Experimental paradigm and hypothesis space of the relation between explicit predictions and ocular expectations. A.** (Left) Experimental setup. In virtual reality, participants sat next to a table and when the target appeared (i.e., butterfly), they reached towards it. (Right) Trial Flow. A colored cue prompted participants to explicitly predict the upcoming target's location (right/left) via the controller's touchpad, with no display change post-response. Following the response, a fixation cross preceded their entry into the virtual environment, where, after 300 milliseconds, a target appeared. Horizontal gaze direction during this 300 milliseconds interval was used to examine ocular expectations. Following the butterfly's appearance participants were instructed to reach towards it as quickly and accurately as possible. **B.** The hypothesis space the relation between explicit predictions and ocular expectations consisted of three hypotheses: (I) the two learning systems are completely distinct with no overlap. (II) they exhibit a partial convergence with some aspects overlapping and others differing (III) There is a near-total overlap, and the two systems fully share information. **C.** Expected results for each hypothesis. (Left) Operationalization of convergence, when both the explicit response and gaze are in the same direction (e.g., "right"), and divergence, when the two are in different directions.

(Right) Expected results for each hypothesis calculated as the percent trials in which they converge or diverge. Note that this is a single example of the analyses used to examine the hypothesis space. See Results for additional analyses.

## Results

### Explicit learning of associative rule

We confirmed that participants exhibit well established signatures of associative learning. Across all experiments accuracy was significantly higher than chance (Mean accuracy= 59.6%, 95% CI= [58.6, 60.6], $t_{110} = 18.8$, $p < .001$, Cohen's $d = 11.1$, see Fig. S1 for results of individual experiments), and was significantly higher on trials in which the cue was valid as opposed to trials with an invalid cue (Mean difference = 44.6%, 95% CI= [40.9, 48.4], $t_{110} = 2.9$, $p < .001$, Cohen's $d = 2.3$, see Fig. S1). Finally, accuracy within blocks exhibited the classic learning curve (see Fig. S2). Thus, participants were engaged and learned the underlying associative rule. Given our primary interest in the learning of the underlying rule and following the pre-registration, in the subsequent analyses accuracy refers to whether the response confirmed to the underlying rule (i.e., '*rule accuracy')*, as opposed to whether the response matched the target's location given that the latter is partially random due to invalid cues.

### Convergence of explicit predictions and ocular expectations

#### Gaze follows explicit predictions and tracks learning processes

Supporting the convergence of explicit predictions and ocular expectations, participants looked significantly more in the direction of their explicit prediction than in the opposite direction while awaiting target appearance. Thus, the difference in gaze direction in trials in which the prediction was "right" as opposed to "left" was significant and large (*Exp. 1*: Mean normalized gaze difference = 0.95, 95% CI= [0.80, 1.10], $t_{24} = 660.77$, $p < .001$, Cohen's $d = 2.56$, *Exp. 2*: Mean normalized gaze difference = 0.90, 95% CI= [0.79, 1.01], $t_{24} = 922.83$, $p < .001$, Cohen's $d = 3.04$, see Fig. 2A). Furthermore, 54 of 56 participants showed a significant difference at the individual level (see SM Fig. S3). Likewise, binarizing gaze direction into an ocular 'response' (i.e.," right"/ "left"), it was

identical to the explicit response in 73% of trials (95% CI = [70%, 75%], see Fig. 2B & Fig. S4). Thus, as expected, ocular expectations robustly align with explicit predictions.

Following the convergence of gaze direction and explicit responses, we examine whether gaze metrics track computationally derived variables underlying explicit learning. Specifically, we focus on Gaze Prediction, a continuous measure of the magnitude of alignment between gaze direction and the explicit prediction, and the underlying probability of the explicit choice made on each trial (i.e., $p(\text{Choice}_t)$). This probability is derived from the well-established family of Rescorla-Wagner (RW) learning models. While we initially planned to use a simple RW model, as per our pre-registration, after comparing models comparison (see Fig. S5) we opted for a RW model that included a stickiness parameter (Culbreth et al., 2016; Daw et al., 2011) (but see similar results with simple RW model in SM ). For each participant we correlated the trial-by-trial $p(\text{Choice}_t)$ and Gaze Prediction, resulting in a Pearson correlations coefficient per participant (see Fig. 2C). In line with our pre-registration, the group distribution of correlation coefficients was significantly greater than zero in both experiments (*Exp. 1*: Mean $\rho = .13$, 95% CI = [.08, .19], $t_{24} = 4.84$, $p < .001$, Cohen's $d = 0.97$, *Exp. 2*: Mean $\rho = .11$, 95% CI = [.05, .16], $t_{24} = 4.02$, $p = .001$, Cohen's $d = 0.82$; Fig. 2D & Fig. S6). Thus, when the underlying probability of the explicit prediction was greater, gaze was more strongly directed in the prediction's direction. Accordingly, Gaze Prediction mirrors fluctuations in the underlying probability of the explicit choice made.

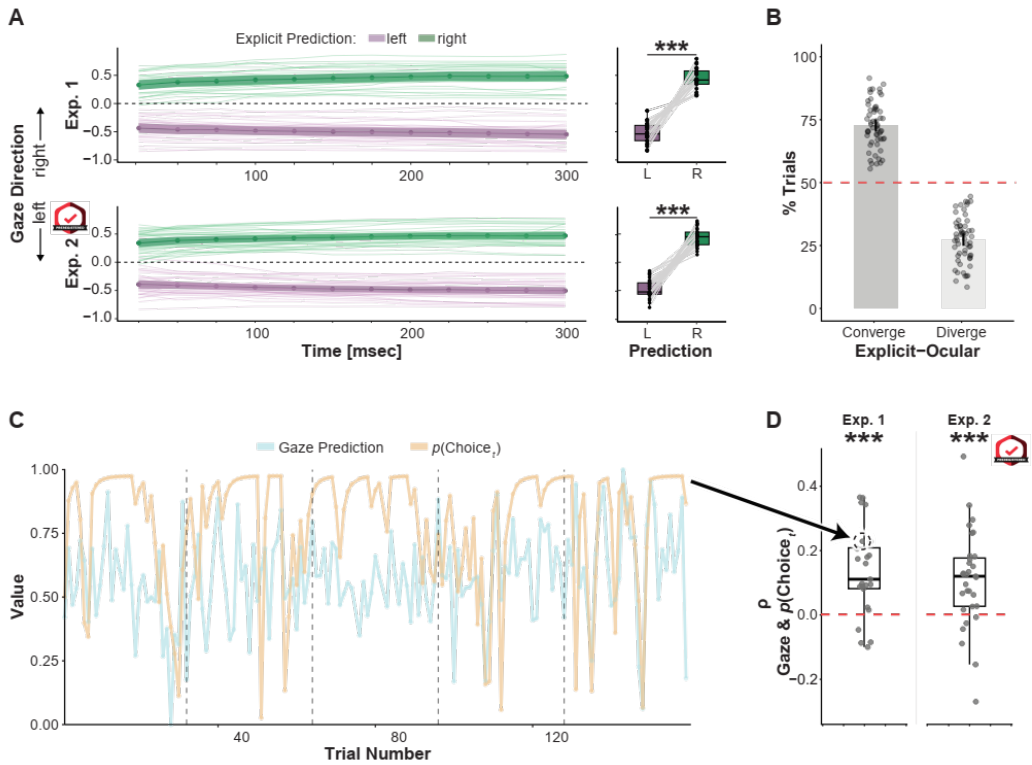


**Figure 2. Convergence of explicit predictions and ocular expectations. A.** Gaze Direction follows explicit prediction. (Left) Time course of normalized gaze direction in 300 msec interval following the explicit prediction and prior to the target's appearance. (Right) Mean normalized gaze direction averaged across interval and trials is robustly aligned with the explicit prediction's direction. **B.** Convergence and divergence of explicit and ocular responses across participants. Gaze was binarized into a ("right"/ "left") 'response' analogous to explicit predictions. Across both experiments, the two types of response were identical in ~70% of the trials. Black line reflects 95% confidence interval, and gray dots are individual participants. **C.** Correlation between trial-by-trial Gaze Prediction magnitude and the latent learning variable $p(\text{Choice}_t)$. A random participant's trajectory across the experiment of $p(\text{Choice}_t)$, the probability of making the explicit choice derived from the computational model, and Gaze Prediction, gaze magnitude normalized by the direction of the explicit response. Note that for visualization's clarity Gaze Prediction values were bound to the limits of $p(\text{Choice}_t)$ between 0-1. **D.** Distribution of individual participants' correlation coefficients between $p(\text{Choice}_t)$ and Gaze Prediction. In both experiments the coefficient's distributions were significantly positive. Thus, as the probability of making a given explicit prediction was greater, gaze was more strongly directed in the prediction's direction.

### Gaze Prediction exhibits computational hallmarks of confidence

Gaze's close relation to $p(\text{Choice}_t)$ resonates with the prevalent conceptualization of confidence as the probability of the choice being correct (De Martino et al., 2013;

Fleming & Daw, 2017; Hertz et al., 2018; Sanders et al., 2016). Accordingly, we tested whether Gaze Prediction exhibits additional characteristics of subjective confidence. Specifically, we focused on Sanders et al., (2016) three computational hallmarks of confidence: (1) a monotonic relation with accuracy, (2) a 'folded X' pattern in which Gaze Prediction increases with a rise in trial evidence for correct trials and decreases for incorrect trials, and (3) trials with a higher Gaze Prediction magnitude showed a stronger relation between accuracy and evidence than trials with low Gaze Prediction magnitude (see Figure 3 & SM for further details of analysis). Fulfilling the first benchmark, we found a positive monotonic relation between accuracy and Gaze Prediction. The group's correlation coefficient distribution was significantly greater than zero (*Exp. 1*: Mean $\rho$ = .33, 95% CI = [.14, .53], $t_{24}$ = 3.5, $p$ = .002, Cohen's $d$ =0.7, *Exp. 2*: Mean $\rho$ = .28, 95% CI = [.11, .45], $t_{24}$ = 3.4, $p$ = .002, Cohen's $d$ =0.61; Fig. 3A & Fig. S7). To test the second benchmark and following our pre-registration, we used trial-by-trial expectation strength ($v$) derived from the learning model as a proxy for trial evidence. In correct trials, Gaze Prediction increased with $v$, whereas for incorrect trials gaze decreased, exhibiting the 'folded X' pattern (see Fig. 3B). The interaction of accuracy and $v$ was significant (*Exp. 1*: $\chi^2_{(1)}$ = 9.32, $p$ = .002, Exp. *2*: $\chi^2_{(1)}$ = 20.67, $p$ < .001; see formula 1 in Table 1 & Table S5 &S6 for pairwise comparisons). Finally, examining the third benchmark, high Gaze Prediction trials exhibited a stronger relation between $v$ and accuracy than low gaze trials (see Fig. 2C), and the interaction was significant (*Exp. 1*: mean (SEM) $\beta$ = 0.15 (0.05), $t_{24}$ =3.16, $p$ = .002; *Exp. 2*: mean (SEM) $\beta$ = 0.16 (0.04), $t_{30}$ =3.60, $p$ < .001; see formula 2 in Table 1). Thus, Gaze Prediction exhibits confidence-like characteristics, fulfilling the three computational benchmarks of subjective confidence.

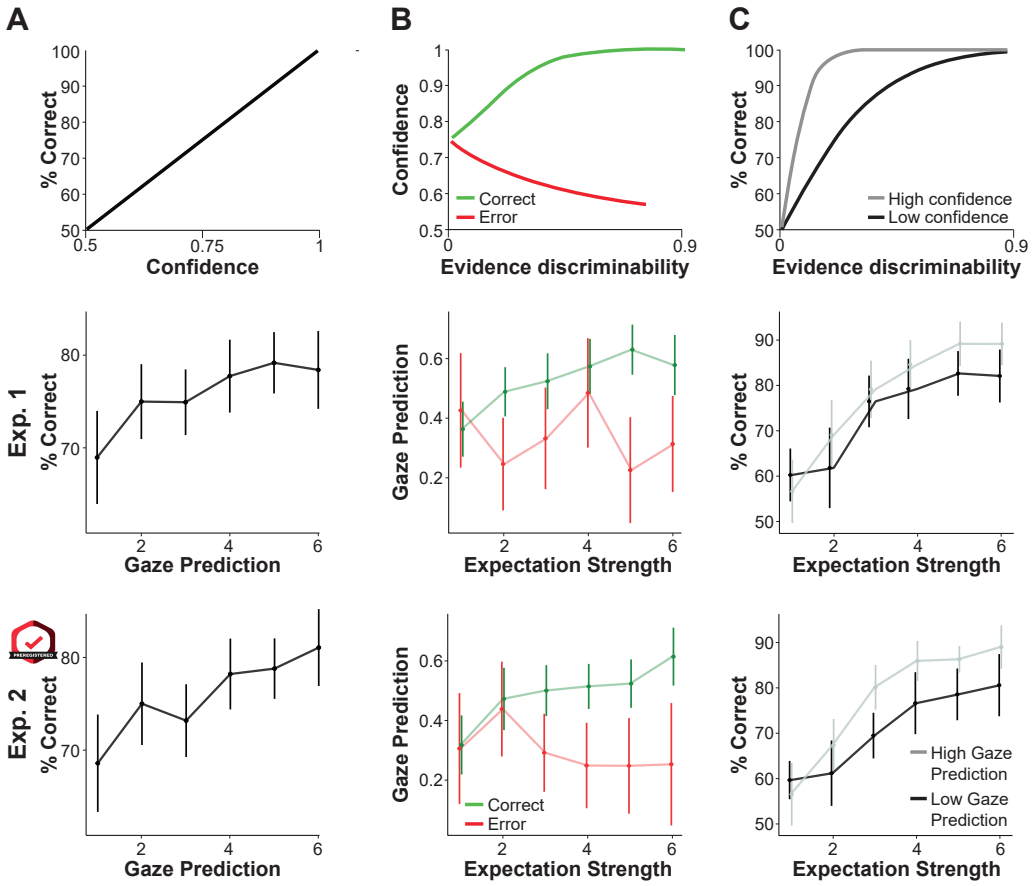


**Figure 3. Gaze Prediction exhibits computational hallmarks of confidence. (Top panel)** Subjective confidence exhibits three computational benchmarks of statistical confidence (figure adapted from Sanders et al., (2016)). (A) A monotonic relation between confidence and accuracy. (B) 'Folded X' pattern, such that as evidence increases confidence in correct trials (green) increases, whereas for incorrect trials (red) it decreases. (C) Steeper slope between evidence strength and accuracy for high confidence trials (grey) as opposed to low confidence trials (black). **(Middle and Bottom panels)** Gaze Prediction, the congruence between gaze's direction and magnitude and explicit prediction, exhibits computational benchmarks of confidence in an exploratory (Exp. 1) and pre-registered replication (Exp. 2) experiment. In these analyses, Gaze Prediction was used in place of subjective confidence, and the expectation strength ($v$) served as a proxy for the trial's evidence.

**Table 1. Regression Models**

| Formula # | Hypothesis | Model Formula |
|---|---|---|
| 1 | Gaze Prediction exhibits 'folded x pattern' (benchmark 2). | Gaze Prediction ~ accuracy*expectation strength + (1\|participant) |

| | | |
|---|---|---|
| 2 | Stronger relation between accuracy and expectation for high Gaze Prediction trials (benchmark 3). | logit(accuracy) ~ Gaze Prediction*expectation strength + (1\|participant) |
| 3 | Explicit confidence is linked to Gaze Prediction beyond other factors. | Gaze Prediction ~ explicit confidence +accuracy +$p$(Choice $_t$) +$\delta_{t-1}$ +(1\| participant) |

**Modest correlation between Gaze Prediction and explicit confidence ratings**

Following Gaze Prediction's confidence-like characteristics, we directly examined the relation between explicit ratings of confidence and gaze. In Experiment 3, after predicting the target's location participants rated their confidence (see SM for full results of pre-registration). In line with previous work (Hertz et al., 2018; Meyniel et al., 2015), trial-by-trial fluctuations in confidence ratings closely tracked the underlying probability of the choice made (i.e., $p$(Choice $_t$)). For each participant we correlated the trail-by-trial $p$(Choice $_t$) and confidence rating, and the group distribution of the correlation coefficients was significant (Mean $\rho$ = .38, 95% CI = [.33, .44], $t_{31}$ = 14.47, $p$ < .001, Cohen's $d$ = 2.56). Furthermore, 30 of the 32 participants exhibited a significant correlation at the individual level. In a similar manner, we examined the relation between Gaze Prediction and explicit confidence. Supporting a significant degree of convergence between Gaze Prediction and confidence, we found that the group level distribution of correlation coefficients was significantly greater than zero (Mean group $\rho$ = .076, 95% CI = [.038, .11], $t_{31}$ = 4.1, $p$ < .001, Cohen's $d$ =0.72). To ensure the correlation between gaze and confidence is not driven by confounding factors (e.g., $p$(Choice $_t$), response accuracy), a linear mixed model incorporating these factors (see Table 1 equation 3) revealed a significant relationship between gaze and confidence ($\chi^2_{(1)}$ = 9.5, $p$ = .002), even after accounting for these factors. Nonetheless, despite the significant association between the two, the correlation magnitude is modest and below our pre-registered criterion ($r$ >. 1) for defining the two processes as similar.

To further explore their partial overlap, we compared the two measures' metacognitive sensitivity by examining the difference in mean confidence on correct and incorrect trials (i.e., Δ confidence, see SM for further justification of this measure). Both measures exhibited significant metacognitive sensitivity (i.e., Δ confidence > 0*; Explicit*

*Δ Confidence*: Mean = 0.58, 95% CI = [.47, .70], $t_{31}$ = 10.44, $p$ < .001, Cohen's $d$ =1.85, *Gaze Δ Confidence*: Mean = 0.10, 95% CI = [.03, .17], $t_{31}$ = 3.01, $p$ = .004, Cohen's $d$ =0.54). Explicit ratings exhibiting significantly higher metacognitive sensitivity than gaze ($t_{31}$ = 7.7, $p$ < .001, Cohen's $d$ =1.36). Interestingly, the two sources' metacognition was not correlated, possibly pointing to distinct mechanisms (see Fig. S8). Thus, Gaze and explicit confidence show a significant, albeit modest, overlap in their contribution to decision making processes, and explicit ratings exhibit enhanced metacognition.

## Divergence of Gaze & Explicit responses

### Explicit responses exhibit superior learning relative to ocular

In approximately 70% of trials the explicit response and ocular 'response' were identical, indicating a robust convergence of the two sources. Notably, in the remaining 30% of trials explicit and ocular responses differed. We turn to examine the consistent divergence of the two learning systems, focusing on their distinct characteristics and respective underlying learning mechanisms. To increase statistical power in the subsequent exploratory analyses, we combined data from the first two experiments which had identical procedures and produced similar results. First, examining explicit and ocular responses' accuracy, both were significantly greater than chance ( *Ocular*: Mean Acc. = 65.2%, 95% CI = [62.2%, 68.2%], $t_{55}$ = 17.52, $p$ < .001, Cohen's $d$ = 2.34; *Explicit*: Mean Acc. = 75.7%, 95% CI = [73.9%, 77.5%], $t_{55}$ = 28.91, $p$ < .001, Cohen's $d$ = 3.86; see Fig. 4c ), and explicit responses were significantly more accurate ($t_{55}$ = 13.13, $p$ < .001, Cohen's $d$ = 1.75). Likewise, explicit responses exhibited a steeper learning curve and higher plateau than ocular (see Fig. 4d). Thus, explicit responses exhibit superior and quicker learning than ocular 'responses'.

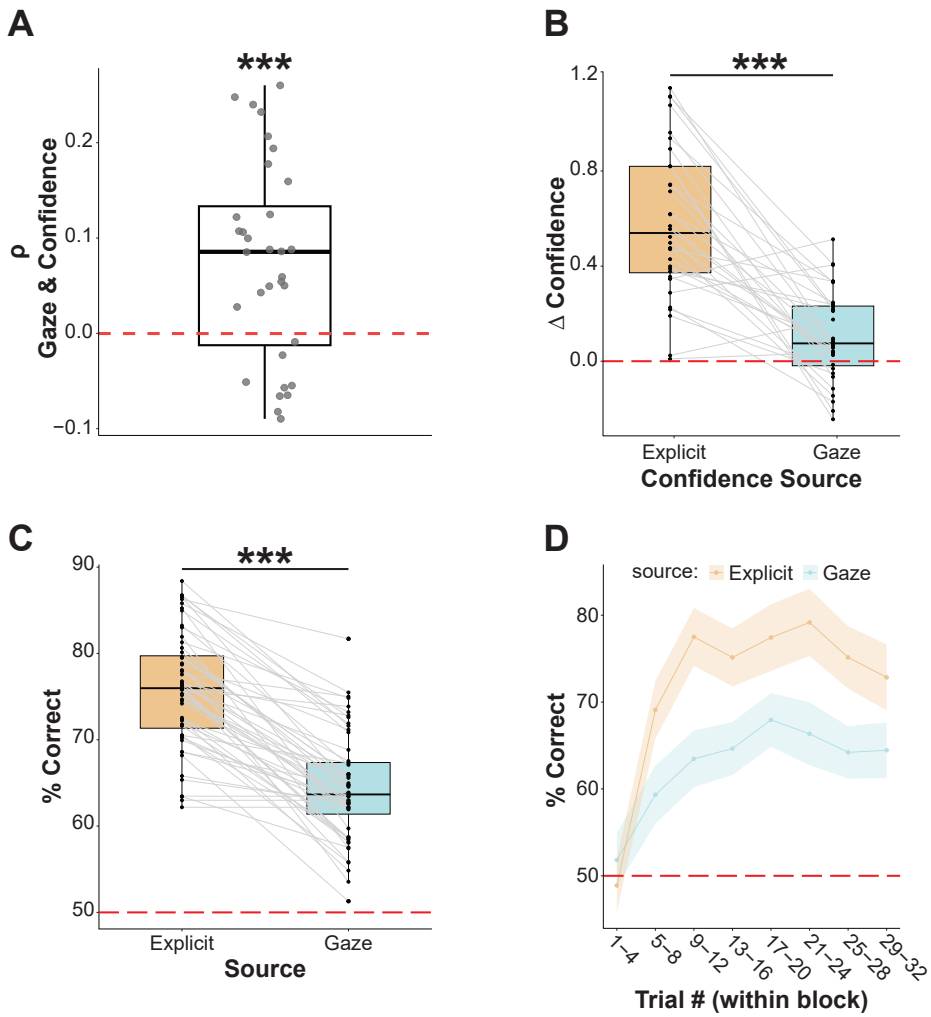


**Figure 4. Divergence of explicit and ocular confidence and responses. (A)** Distribution of individual participants' correlation coefficients between explicit confidence ratings and gaze derived confidence (i.e., Gaze Prediction) in Exp. 3. Mean correlation was 0.07 and was significantly greater than zero ($p < .001$), yet its magnitude was modest and smaller than our pre-registered criterion ($r >$. 1) for defining the two processes as similar. **(B)** Comparison of explicit and gaze derived confidences' metacognitive sensitivity (i.e., Δ confidence). Explicit ratings were significantly more sensitive than gaze ($p < .001$). Importantly both sources of confidence exhibited significant sensitivity and their Δ confidence was greater than chance (dashed red line; both $p$'s < .001). **(C)** Comparison of explicit and ocular 'response' accuracy. Explicit responses were significantly more accurate than gaze ($p < .001$). Both sources' accuracy was significantly greater than chance (dashed red line; both $p$'s < .001). **(D)** Learning curve for explicit and ocular responses. Learning curve was calculated across blocks of consecutive trials governed by the same underlying rule. Shaded areas are 95% confidence intervals across participants. Explicit responses exhibited a steeper learning curve and higher plateau than ocular.

**Ocular learning relies on distinct computational decision mechanism**

To elucidate explicit responses' increased accuracy relative to ocular, we consider two possibilities: (1) ocular responses rely on a noisier learning mechanism that results in degraded learning relative to explicit or (2) ocular learning utilizes a different computational mechanism with distinct characteristics. To assess these possibilities, we examine the learning dynamics and computationally model the trial-by-trial decision process for both types of responses. First, we compare the model agnostic measure of response rule switching. Importantly, the underlying rule was cue contingent (i.e., if green cue signifies "right" then purple cue necessarily is "left"), and participants were informed of this. A rule switch occurs when a participant changes from responding according to Rule A (e.g., green cue with a "left" response or purple cue with a "right" response) on one trial to responding according to Rule B (e.g., green cue with a "right" response or purple cue with a "left" response) on the following trial. Rule switches presumably reflect an updating of the learned model of the environment or its exploration, resulting in a switch. Importantly, they can occur regardless of whether the same cue is presented consecutively and may occur even if the same response (e.g., "right") was made consecutively. Interestingly, ocular 'responses' showed significantly higher rule switching than explicit (*Ocular:* Mean rate = 40%, 95% CI = [38%, 42%], *Explicit:* Mean rate = 21%, 95% CI = [18%, 23%], $t_{55} = 16.34$, $p < .001$, Cohen's $d$ =2.18; see Figure 5a). Control analyses (see SM and Fig. S9) demonstrated that this difference in rule switching cannot be explained by a difference in motor switching (i.e., responding "right" after a "left" response regardless of the cues presented), nor a difference in the effect of the previous trial's accuracy on rule switching. Hence, explicit responses are more likely to preserve their response rule across trials, whereas ocular responses exhibit an increased tendency to switch their rule.

To further understand the systems' learning dynamics, we computationally modeled ocular 'responses' in an identical manner to explicit responses. The model space consisted of: (1) the well-established Rescorla-Wagner (RW) model (Rescorla, 1972), (2)

a Rescorla-Wagner model with an additional stickiness parameter (sRW) (Culbreth et al., 2016; Daw et al., 2011), (3) and a noisy Win-Stay Lose-Switch (WSLS) model (Wilson & Collins, 2019);see Methods and SM for further details of modeling procedures). Importantly, explicit and ocular learning diverged in their selected model assessed via their Bayesian Information Criteria (BIC)(Schwarz, 1978). Explicit learning was best explained by an extended RW with a 'stickiness' parameter (sRW) ($\Delta$ BIC ~ 7, $F_{1.24,68.02}$ = 112.21, $p < .001$, $\eta^2_p = .67$), whereas ocular learning was best explained by the basic RW model that does not explicitly model rule preservation via a 'stickiness' parameter ($\Delta$ BIC ~ 6, $F_{1.43,78.59} = 9.66$, $p = .008$, $\eta^2_p = .15$; see Fig. 5b & Fig. S5). Thus, both types of learning are explained by similar models that share the same learning mechanism and differ in their decision mechanism.

Next, we compare explicit and ocular's sRW model parameters. Their learning rate ($\alpha$) did not significantly differ (see Figure 5c), and a Bayesian analysis found moderate evidence for a lack of difference (*Ocular:* Mean $\alpha = .33$, .95% CI = [.25, .41], *Explicit:* Mean $\alpha = .31$, .95% CI = [.26, .36], $t_{55} = -0.42$, $p = 0.68$, $BF_{01} = 6.31$). Thus, explicit and ocular responses update the perceived value and learn similarly. Further supporting a shared mechanism, explicit and ocular's learning rate was significantly correlated across participants ($r_{55} = 0.33$, $p = .01$; see Figure S11). In contrast, the decision mechanism was significantly different between sources. The stickiness parameter ($\rho$), was significantly higher for explicit responses than ocular responses (*Ocular:* Mean $\rho = 0.21$, .95% CI = [0.15, 0.27], *Explicit:* Mean $\rho = 1.43$, .95% CI = [0.59, 2.27], $t_{55} = 2.29$, $p = 0.005$, Cohen's $d = 0.38$), reflecting explicit responses' increased tendency for rule perseverance. Furthermore, across participants the explicit and ocular $\rho$ was not significantly correlated and there was anecdotal evidence for the lack of a correlation ($r_{55} = 0.06$, $p = .69$, $BF_{01}$=3.1; see Figure S11). Comparing the sources' inverse temperature ($\beta$) yielded inconclusive results. Although, explicit responses had higher $\beta$ values than ocular (*Ocular:* Mean $\beta = 9.11$, .95% CI = [2.21, 14.02], *Explicit:* Mean $\beta = 12.45$, .95% CI = [3.6, 21.28]), indicating that explicit responses were more deterministic and followed the perceived value ($v$) of the choice. Statistically testing this difference, a paired t-test did not find a significant difference

between the two ($t_{55}$ = -0.65, $p$ = 0.52; though see SM for further analyses and Fig. S10). Across participants the explicit and ocular β was not significantly correlated and there was anecdotal evidence for the lack of a correlation ($r_{55}$ = -0.04, $p$ = .76, $BF_{01}$=3.2; see Figure S11). Thus, ocular 'responses' may be more exploratory and stochastic, although this finding must be interpreted cautiously. To summarize, explicit and ocular learning dynamics attribute values to choices in an identical manner and differ in manner that they incorporate the perceived value into the decision process. Specifically, explicit responses are 'stickier' than ocular responses, tending to adhere to the rule previously learned, while ocular responses are more exploratory.

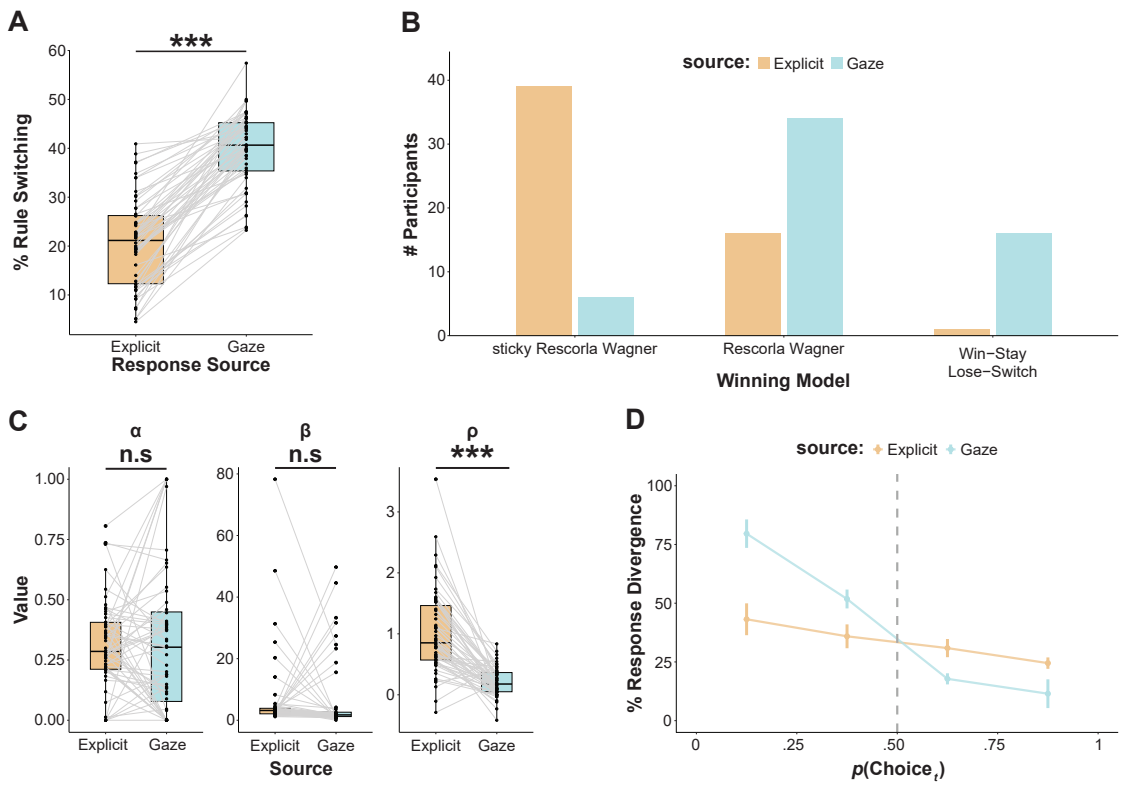


**Figure 5. Distinct learning mechanisms of explicit and ocular learning. (A)** Ocular responses exhibit significantly more rule switching than explicit responses that tend to preserve in their response rule across trial. **(B)** Distribution of winning computational model for ocular and explicit responses across participants. Explicit learning dynamics were mainly explained by a Rescorla-Wagner model that included a 'stickiness' parameter (sRW) that accounts for the tendency to persevere to the previous trial's response rule regardless of its outcome. In contrast ocular learning dynamics were better explained by a simple RW model or Win-Stay-Lose-Switch model that doesn't incorporate a tendency for preservation per se. **(C)** Comparison of explicit and ocular

sRW parameters. Model parameters include learning rate (α), inverse temperature (β) that reflects decision stochasticity, and a perseveration parameter (ρ). Only ρ was significantly greater for explicit responses, whereas β showed inconclusive results. Note that in β panel we only included values whose deviation was < 3 S.D. from group mean to enable enhanced resolution of visualization (see SM for further discussion and visualizations). **(D)** Explicit and ocular responses were more likely to differ in exploratory trials in which the response made had a lower probability (i.e., $p(\text{Choice}_t)$) that was calculated based on explicit (brown) or ocular (turquoise) response trajectories. Both showed a similar effect that in exploratory trials (i.e., $p(\text{Choice}_t) < .5$) the two responses were more likely to differ from one another in contrast to exploitation trials (i.e., $p(\text{Choice}_t) > .5$).

## Ocular and explicit responses diverge on exploratory trials

Exploring the interaction of ocular and explicit learning models, we investigate the potential benefits of their convergence and divergence. Supporting the beneficial impact of their alignment, converging trials in which the ocular and explicit responses were identical, showed significantly higher explicit accuracy than diverging trials (*Converging*: Mean Acc. = 78%, .95% CI = [76%, 79%], *Diverging*: Mean Acc. = 71%, .95% CI = [68%, 74%], $t_{55} = 4.75$, $p < 0.001$, Cohen's $d = 0.63$, see Figure S12). Thus, the convergence of the learning models reflects improved learning. Turning to examine in which trials responses diverge, we found that ocular and explicit responses show significantly higher divergence in trials that are exploratory and with a lower $p(\text{Choice}_t)$ (i.e., <.5; *Exploratory*: % Divergence = 35%, .95% CI = [31%, 40%], *Exploitation*: % Divergence = 26%, .95% CI = [23%, 28%], $t_{55} = 4.13$, $p = 0.001$, Cohen's $d = 0.56$, see SM & Figure S13). Thus, divergence of responses may reflect a 'hedging of one's bet' in trials in which the choice made does not reflect the current evidence.

## Discussion

The study examined the interplay between explicit predictions and implicit oculomotor expectations in associative learning, revealing a partial overlap in their learning processes. On the one hand there was robust convergence between explicit predictions of the upcoming target's location and anticipatory gaze dynamics. Furthermore, gaze exhibited the computational hallmarks of confidence, reflecting an oculomotor confidence-like assessment of one's explicit prediction. On the other hand,

the two substantially diverged in several central aspects. Explicit learning exhibited significantly increased accuracy, a steeper learning curve, and enhanced metacognitive sensitivity relative to ocular learning. Examining their underlying learning mechanisms, the two types of responses were best described by similar yet different models. Specifically, both utilized the same learning mechanism, but different decision mechanisms. Explicit predictions exhibited perseverance of the learned rule, while gaze was less affected by this consideration. Finally, a possible benefit of two partially overlapping learning processes, relates to their increased divergence on exploratory trials that enables the hedging of one's 'bet' under uncertainty. These findings, underscore the complex interplay between the learning manifested by an explicit response system and implicit oculomotor system, that together enable embodied agents to learn adaptively in an uncertain environment.

**Convergence of explicit predictions and ocular expectations**

Explicit predictions and ocular expectations converged robustly across several dimensions. Gaze direction was identical to the explicit prediction of the upcoming target's location in approximately two-thirds of the trials. Furthermore, trial by trial fluctuations in Gaze Prediction, a measure of the alignment between gaze and the explicit predication, tracked the computationally derived likelihood of the explicit choice made. Following Posner's (1980) work on spatial attention, a straightforward explanation for this convergence is that explicit responses guide attention to the expected location and anticipatory gaze reflects a motor preparation process that enables a quicker response when the target appears. Similarly, in value-based decisions, pre-decision gaze patterns have been robustly linked to attention and the value ascribed to an item prior to the explicit choice between competing items (Krajbich et al., 2010; Krajbich & Rangel, 2011; Thomas et al., 2019). Furthermore, recent work in primates found that gaze towards the location of a previously presented target, tracks the reevaluation and neural encoding of the item's value (Ferro et al., 2024). Thus, gaze dynamics are closely linked to attention and latent computational processes that lead up to an explicit decision. The

current study expands these findings by examining post-decision gaze dynamics during associative learning where the value is constantly being learned and is not fixed a priori.

An additional explanation for the robust convergence is that explicit predictions induce an implicit representation of the expected location, resulting in a gaze shift that is not task dependent. Recent work demonstrated that activation of stimuli in working memory results in an involuntarily gaze shift to the stimulus' location even when location is task irrelevant (Van Ede, Chekroud, Stokes, et al., 2019; van Ede et al., 2020). Similarly, during the use of abstract concepts such as a 'mental number line' gaze patterns reflect the concept's spatial organization (Viganò et al., 2024). In the current study, we cannot distinguish between these complementary explanations since gaze always followed the participant's response, leaving open the question of whether gaze dynamics exhibits learning during passive viewing without an explicit response. Despite the impressive convergence between explicit predictions and ocular expectations, they diverge in about a third of the trials, highlighting putative differences between these systems and their partial overlap.

**Gaze reflects a second-order assessment of explicit predictions**

Their partial overlap indicates that while gaze dynamics track explicit decisions, they may also carry additional information partially diverging from the explicit prediction. Confidence, a second-order assessment of the accuracy of one's judgment, is a prime example of an additional facet of the decision process that partially overlaps with the initial explicit decision. In the current study, the post-decision Gaze Prediction exhibited the computational hallmarks of confidence (Sanders et al., 2016). Whether these hallmarks are sufficient to equate gaze with sensorimotor confidence resonates with similar debates regarding other measures such as neural activation and animal behavior that exhibit these hallmarks (Bang & Fleming, 2018; Rausch & Zehetleitner, 2019; Schmack et al., 2021). Following the significant albeit modest correlation between explicit confidence and Gaze Prediction in experiment 3, we argue that gaze exhibits *confidence-like* characteristics and reflects a parallel second-order sensorimotor assessment that resembles explicit confidence yet should not be equated with it.

Importantly, both explicit confidence and Gaze Prediction closely tracked the likelihood of the choice made (i.e., $p(\text{Choice}_t)$), and this may be a core feature of second-order judgments (Hertz et al., 2018; Meyniel et al., 2015).

Understanding the relation between explicit and ocular second-order assessments touches upon several issues in the field of confidence and metacognition. First, dissociating their unique biases may further shed light on different types of metacognitive inefficiency (Shekhar & Rahnev, 2021) that prevail in different second-order processes. Given that ocular responses exhibit reduced rule perseverance across trials, we speculate that ocular confidence may exhibit reduced serial dependence effects relative to serial dependence effects found in explicit confidence (Mei et al., 2023). Second, our findings relate to ongoing debates whether metacognition is a unitary domain-general or domain-specific capacity (Arbuzova et al., 2023; Rouault et al., 2018). The lack of correlation between explicit and ocular metacognition made on the same decision is more easily interpreted within a domain-specific framework in which there is no assumed relation between different types of metacognitions. These findings should be interpreted with caution given the small sample size and that the addition of an explicit rating of confidence in Exp. 3 slightly altered gaze dynamics and weakened the relation between gaze and confidence. Despite these limitations, an intriguing possibility is that the two second order judgments are not correlated yet their integration (e.g., the average of explicit and ocular confidence) may enable enhanced metacognition as opposed to the use of a single measure. This possibility underscores the importance of understanding diverging cognitive processes within the broader context of parallel processes whose integration enhances adaptive behavior.

**Divergence of explicit and ocular decision mechanisms**

Explicit predictions and ocular expectations also diverged in several key aspects. Modeling their learning mechanisms, both utilized the same value-assignment mechanism, and diverged in their decision mechanisms. Presumably this enables a compact learning process with similar value ascriptions across systems, allowing adaptive behavior based on the specific environment and systems' goals. In contrast,

other instances, such as learning semantic and affective representations, there is evidence for parallel learning processes (Heimer et al., 2023). Thus, the manner systems are integrated differs based on the systems involved and task requirements. Understanding how systems interact is especially relevant in understanding psychopathology in which this integration may be disrupted. This is especially relevant for schizophrenia spectrum disorders in which an aberrant interaction between an explicit and implicit sensorimotor systems is postulated to be a core deficit (Krugwasser et al., 2021; Nelson et al., 2014; Stern et al., 2020).

Examining their decision mechanism, explicit responses exhibited increased rule perseverance, whereas ocular responses did not tend to persevere. Motor perseverance (e.g., repeating button presses across trials) is associated with model-free learning even when it is task irrelevant (Shahar et al., 2019). In contrast, here we found that explicit responses exhibit response-rule perseverance, tending to stick to the learned rule, regardless of the motor response. One explanation for this difference is that explicit responses reflect higher-order processes that aspire for self-consistency across trials (Koriat, 2012). In contrast, ocular responses reflect a more present-focused decision. This is in line with Gaze Prediction's confidence-like characteristics that assess the current trial's accuracy. An alternative explanation is related to the utility of stickiness in a static environment that enables rapid learning (Daw et al., 2006; Findling et al., 2019). Importantly, at the start of the experiment participants were informed that the environment is static with a fixed ratio (i.e., 3:1) of valid and invalid trials. It may be that explicit responses can utilize this 'meta-knowledge' of the environment and adapt an adaptive sticky decision mechanism, whereas ocular responses may be unable to utilize this knowledge and require learning through experience, resonating with the distinction between model-free and model-based learning (Daw et al., 2006; Decker et al., 2016; Dolan & Dayan, 2013). Future work that examines explicit and ocular learning in a dynamic environment may resolve between these different explanations.

Conceptualizing cognitive capacities as multimodal processes that spans across systems is critical in understanding embodied human cognition that encompasses both

explicit decisions as well as a multitude of sensorimotor systems. Indeed, the oculomotor behavior examined in the current study is a single sensorimotor system that exhibited a robust link to explicit decisions. Previous research highlights the contribution of other sensorimotor systems such as pupillometry (Finke et al., 2021; Preuschoff et al., 2011) and motor kinematics (Rolfs & Schweitzer, 2022; Shadmehr et al., 2019). Understanding the interaction and integration of explicit and sensorimotor systems is critical in developing a more ecological and embodied understanding of human behavior.

## Materials and Methods

We conducted three experiments. An initial exploratory Experiment 1, and two pre-registered experiments (Experiment 2: https://osf.io/p4hyq, 09.05.2023; Experiment 3: https://osf.io/ngsx6. 17.07.2023). The pre-registration analysis plan was followed, unless otherwise stated. For consistency, we retrospectively implemented the pre-registered analysis plan for Experiment 1. The primary inclusion criteria were: (1) Explicit response accuracy (i.e., the response matches the target's location) above 55%. This value was chosen to ensure that parameter fitting was reliable and that participants understood the task. (2) the VR and eye-tracking recording was valid in at least 70% of trials.

### Participants

For Experiment 1 we collected data from 35 participants, and 25 participants (19 female, mean age 23.9, SD 4 years) were included in analysis. For Experiment 2, we pre-registered 40 participants, anticipating that ~30 would provide sufficient statistical power to replicate Experiment 1's effects (see https://osf.io/p4hyq for power analysis). The final sample included data from 31 participants (21 female, mean age 23.7, SD 3.4 years). For Experiment 3, we pre-registered 40 participants, and the final sample size included data from 32 participants (23 female, mean age 25.3, SD 5.3 years).

All participants were right-handed, with normal or corrected-to-normal vision and self-reported no-history of neurological impairments and were recruited via social media. All participants gave written informed consent and were compensated (50 NIS per hour or

course credits) for their participation. Experiment 1 and 2 were performed at Bar-Ilan University and experiment 3 at Haifa University. The universities' Institutional Review Board approved this study (Approval Numbers: ISU202107003 & HU448/19 for Exp. 1&2 and Exp. 3 respectively).

**Hardware and Setup**

Participants wore an HTC Vive Pro Eye head-mounted display connected to a high-performance desktop, which tracked their hand movements and projected a virtual environment created in Unity (2018.3.2) at a 90Hz refresh rate. Responses were recorded through the controller's trackpad, and eye movements were monitored using the built-in Tobii trackers (~90Hz; see Applebaum et al., (2024) and SM for additional technical details).

**Experimental Procedure**

We adapted a well-established associative learning task that has been previously modeled (Frässle et al., 2021; Harrison et al., 2021; Iglesias et al., 2013) to VR. Over 160 trials, participants were presented with one of two cues (i.e., a colored square), and they predicted, via a button press on the controller, the upcoming target's location ("right"/ "left"). Following their prediction and a fixation cross (duration 1.5 seconds), participants entered a virtual environment and the target, a butterfly, appeared after a brief delay (300 msec). Once they found the target, participants were instructed to reach towards it, and a virtual hand tracked the actual hand movement (see Figure 1A for trial flow). Participants' goal was to learn the underlying rule governing the cue-location association and respond accordingly. Participants were informed that: (1) in 75% of the trials, the cue is Valid, following an underlying rule, and in the remaining trials the cue is Invalid. (2) The underlying rule is contingent between the two cues (e.g., if cue A indicates a right target, cue B necessarily indicates left) (3) Throughout the experiment, the underlying rule can be switched numerous times, requiring participants to detect these changes. The rule was changed five times during the experiment, with each block lasting between 29 and 35 trials (see SM for further details). In Experiment 3 participants performed an identical task, with the addition that following their prediction of the target's location

they rated their confidence on a five-point vertical slide. Following the associative learning task, participants also performed a second experiment with a sensorimotor conflict detection task to assess their Sense of Agency (Applebaum et al., 2024), and completed several self-report questionnaires pertaining to psychotic-like and anomalous self-experiences (Cicero et al., 2017; Foa et al., 2002; Loewy et al., 2011), whose results are beyond the scope of the current work.

**Measures**

In each trial, we analyzed participants' explicit predictions (target will appear in Right/ Left location) and their implicit ocular expectation of the target's location. Predictions were assessed for *actual accuracy* (whether the target appeared as predicted) and *rule accuracy* (conformity to the underlying rule). Our primary focus was on rule accuracy due to our interest in the learning of the underlying rule, except when noted otherwise.

Ocular expectations were derived from a pre-registered, 300 msec interval in the VR environment, after participants made their explicit prediction but critically before the target appeared. We normalized and averaged horizontal gaze directions across this interval for each trial, yielding a single Gaze Direction value per trial. From Gaze Direction we derived: (1) a binary *Gaze Response* (“Right" or "Left"), that is analogous to the explicit prediction. (2) *Gaze Prediction* a continuous measure obtained by normalizing Gaze Direction based on the trial's explicit prediction. This measure indicates the alignment between explicit prediction and gaze direction, with positive values denoting gaze towards the predicted direction and negative values indicating gaze in the opposite direction (see SM for transformation details).

**Modeling**

To analyze learning dynamics throughout the experiment, we employed a model-based approach. The model space consisted of: (1) the well-established Rescorla-Wagner (RW) model (Rescorla, 1972), (2) a Rescorla-Wagner model with an additional stickiness parameter (sRW) (Culbreth et al., 2016; Daw et al., 2011), (3) and a noisy Win-Stay Lose-Switch (WSLS) model (Wilson & Collins, 2019).

The RW model consists of a perceptual and decision model. In the perceptual model each option's value ($v$) is updated the prediction error (δ)—the discrepancy between the previous trial's expected and actual outcome ($r$)—modulated by the learning rate ($\alpha$), a free parameter that controls the extent of updating from past trials (Equation 1A). Lower $\alpha$ values suggest slower learning over more trials. The value ($v$) from the perceptual model informs the decision model, which employs a SoftMax function to calculate the choice probability (i.e., $p$(Choice $_t$)). The function's steepness and randomness are dictated by the inverse temperature (β), another free parameter, reflecting decision consistency (Equation 1B).

$$v_{t+1} = v_t + \alpha * \delta_t$$

$$\delta_t = r_t - v_t$$

(Equation 1A)

$$\boldsymbol{P}._{Choice} = \frac{e^{\beta * \boldsymbol{v}}}{e^{\beta * \boldsymbol{v}} + e^{\beta * (1-\boldsymbol{v})}}$$

(Equation 1B)

The "sticky" Rescorla-Wagner (sRW) model extends the traditional model by adding a perseveration parameter (ρ) to quantify the tendency to repeat the previous trial's decision (Culbreth et al., 2016; Daw et al., 2011). This parameter is included in the decision model (Equation 2), multiplied by *rep (a)*, an indicator function that equals 1 for decision repetition and 0 otherwise. Importantly, responses are coded in a cue-contingent manner (cue A & a "right" response are coded identically to cue B & a "left" response), meaning perseveration reflects rule adherence rather than mere motor repetition. Notably, the RW model is nested within the sRW model when the parameter ρ is set to zero.

$$P._{Choice} = \frac{e^{\beta * \boldsymbol{v} + \rho * rep(\mathrm{a})}}{e^{\beta * \boldsymbol{v} + \rho * rep(\mathrm{a})} + e^{\beta * (1-\boldsymbol{v}) + \rho * rep(\mathrm{a})}}$$

(Equation 2)

Unlike the Rescorla-Wagner model, which uses a free parameter to determine the integration window, the WSLS model (Equation 3) bases decisions solely on the previous trial's outcome, incorporating a free parameter ($\boldsymbol{\varepsilon}$) to incorporate probabilistic noise in the decision-making.

$$P._{Choice} \begin{cases} 1 - \varepsilon \;\; if \; C_{t-1} \; correct \\ \varepsilon \;\; if \; C_{t-1} \; incorrect; \end{cases}$$

(Equation 3)

For all models, stimuli and responses were encoded within a unified contingency space to minimize free parameter estimates. This is in line with our pre-registration and previous work (Harrison et al., 2021; Iglesias et al., 2013). Model parameters were estimated via the Bayesian adaptive direct search (BADS) toolbox (Acerbi & Ji, 2017), optimizing each model's log likelihood. Parameter recovery across models was reliable (see Figure S19). Model comparison was performed via the Bayesian Information Criteria (BIC; Schwarz, 1978). All models showed good model recovery (see Fig. S20).

Linear mixed models were fitted using the 'lme4' package (Bates et al., 2014), using likelihood ratio tests to compare the full model against a version without the targeted effect to obtain p-values.

Pre-registrations, analysis code and notebooks available at:

https://osf.io/qtg6k/

Data available at:

https://osf.io/6nj8u/?view_only=0852199aacb94825a3ef81c760b972a6

**References**:


Altmann, G. T. (2004). Language-mediated eye movements in the absence of a visual world: The 'blank screen paradigm.' *Cognition*, *93*(2), B79–B87.

Applebaum, A., Netzer, O., Stern, Y., Zvilichovsky, Y., Mashiah, O., & Salomon, R. (2024). The Body Knows Better: Sensorimotor signals reveal" Suboptimal" inference of the Sense of Agency in the human mind. *bioRxiv*, 2024–02.

Arbuzova, P., Maurer, L. K., & Filevich, E. (2023). Metacognitive domains are not aligned along a dimension of internal-external information source. *Psychonomic Bulletin & Review*, *30*(3), 1125–1135.

Balsdon, T., Wyart, V., & Mamassian, P. (2020). Confidence controls perceptual evidence accumulation. *Nature Communications*, *11*(1), 1753.

Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, *115*(23), 6082–6087.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv Preprint arXiv:1406.5823*.

Biederman, I., Glass, A. L., & Stacy, E. W. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology*, *97*(1), 22.

Chalk, M., Seitz, A. R., & Seriès, P. (2010). Rapidly learned stimulus expectations alter perception of motion. *Journal of Vision*, *10*(8), 2–2.

Cicero, D. C., Neis, A. M., Klaunig, M. J., & Trask, C. L. (2017). The Inventory of Psychotic-Like Anomalous Self-Experiences (IPASE): Development and validation. *Psychological Assessment*, *29*(1), 13–25. https://doi.org/10.1037/pas0000304

Colizoli, O., de Gee, J. W., Urai, A. E., & Donner, T. H. (2018). Task-evoked pupil responses reflect internal belief states. *Scientific Reports*, *8*(1), 13702.

Constant, M., Pereira, M., Faivre, N., & Filevich, E. (2023). Prior information differentially affects discrimination decisions and subjective confidence reports. *Nature Communications*, *14*(1), 5473.

Corbetta, M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Ollinger, J. M., Drury, H. A., Linenweber, M. R., Petersen, S. E., Raichle, M. E., & Van Essen, D. C. (1998). A common network of functional areas for attention and eye movements. *Neuron*, *21*(4), 761–773.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215.

Culbreth, A. J., Westbrook, A., Daw, N. D., Botvinick, M., & Barch, D. M. (2016). Reduced model-based decision-making in schizophrenia. *Journal of Abnormal Psychology*, *125*(6), 777.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110.

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological Science*, *27*(6), 848–858.

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science*, *29*(5), 761–778.
Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1837.
Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.
Dotan, D., Meyniel, F., & Dehaene, S. (2018). On-line confidence monitoring during decision making. *Cognition*, *171*, 112–121. https://doi.org/10.1016/j.cognition.2017.11.001
Drugowitsch, J., Mendonça, A. G., Mainen, Z. F., & Pouget, A. (2019). Learning optimal decisions with confidence. *Proceedings of the National Academy of Sciences*, *116*(49), 24872–24880.
Faivre, N., Vuillaume, L., Bernasconi, F., Salomon, R., Blanke, O., & Cleeremans, A. (2020). Sensorimotor conflicts alter metacognitive and action monitoring. *Cortex*, *124*, 224–234.
Ferro, D., Cash-Padgett, T., Wang, M. Z., Hayden, B. Y., & Moreno-Bote, R. (2024). Gaze-centered gating, reactivation, and reevaluation of economic value in orbitofrontal cortex. *Nature Communications*, *15*(1), 6163.
Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2019). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature Neuroscience*, *22*(12), 2066–2077.
Finke, J. B., Roesmann, K., Stalder, T., & Klucken, T. (2021). Pupil dilation as an index of Pavlovian conditioning. A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *130*, 351–368.
Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, *75*, 241–268.
Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114. https://doi.org/10.1037/rev0000045
Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443.
Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The Obsessive-Compulsive Inventory: Development and validation of a short version. *Psychological Assessment*, *14*(4), 485–496. https://doi.org/10.1037/1040-3590.14.4.485
Frässle, S., Aponte, E. A., Bollmann, S., Brodersen, K. H., Do, C. T., Harrison, O. K., Harrison, S. J., Heinzle, J., Iglesias, S., & Kasper, L. (2021). TAPAS: an open-source software package for translational neuromodeling and computational psychiatry. *Frontiers in Psychiatry*, *12*, 680811.
Gilboa, A., & Marlatte, H. (2017). Neurobiology of Schemas and Schema-Mediated Memory. *Trends in Cognitive Sciences*, *21*(8), 618–631. https://doi.org/10.1016/j.tics.2017.04.013
Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932.
Goupil, L., & Aucouturier, J.-J. (2021). Distinct signatures of subjective confidence and objective accuracy in speech prosody. *Cognition*, *212*, 104661.
Goupil, L., Ponsot, E., Richardson, D., Reyes, G., & Aucouturier, J.-J. (2021). Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nature Communications*, *12*(1), 861.
Harrison, O. K., Köchli, L., Marino, S., Luechinger, R., Hennel, F., Brand, K., Hess, A. J., Frässle, S., Iglesias, S., & Vinckier, F. (2021). Interoception of breathing and its relationship with anxiety. *Neuron*, *109*(24), 4080–4093.

Heimer, O., Kron, A., & Hertz, U. (2023). Temporal dynamics of the semantic versus affective representations of valence during reversal learning. *Cognition*, *236*, 105423.
Helmholtz, H. (1924). Treatise on physiological optics (Vol. 3). *Menasha, WI: Optical Society of America*.
Hertz, U., Bahrami, B., & Keramati, M. (2018). Stochastic satisficing account of confidence in uncertain value-based decisions. *PLoS One*, *13*(4), e0195399.
Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E. M., & Stephan, K. E. (2013). Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. *Neuron*, *80*(2), 519–530. https://doi.org/10.1016/j.neuron.2013.09.009
Johansson, R., & Johansson, M. (2014). Look here, eye movements play a functional role in memory retrieval. *Psychological Science*, *25*(1), 236–242.
Kahneman, D. (2017). *Thinking, fast and slow*.
Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80.
Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298.
Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857.
Krauzlis, R. J., Lovejoy, L. P., & Zénon, A. (2013). Superior colliculus and visual spatial attention. *Annual Review of Neuroscience*, *36*, 165–182.
Krugwasser, A. R., Stern, Y., Faivre, N., Harel, E. V., & Salomon, R. (2021). *Impaired Sense of Agency and Associated Confidence in Psychosis*.
Kustov, A. A., & Lee Robinson, D. (1996). Shared neural control of attentional shifts and eye movements. *Nature*, *384*(6604), 74–77.
Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, *93*(2), 451–463.
Loewy, R. L., Pearson, R., Vinogradov, S., Bearden, C. E., & Cannon, T. D. (2011). Psychosis risk screening with the Prodromal Questionnaire—Brief Version (PQ-B). *Schizophrenia Research*, *129*(1), 42–46. https://doi.org/10.1016/j.schres.2011.03.029
Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021
Maresch, J., Mudrik, L., & Donchin, O. (2021). Measures of explicit and implicit in motor learning: What we know and what we don't. *Neuroscience & Biobehavioral Reviews*, *128*, 558–568.
Mei, N., Rahnev, D., & Soto, D. (2023). Using serial dependence to predict confidence across observers and cognitive domains. *Psychonomic Bulletin & Review*, *30*(4), 1596–1608.
Metcalfe, J., & Shimamura, A. P. (1994). *Metacognition: Knowing about knowing*. MIT press.
Meyniel, F., Schlunegger, D., & Dehaene, S. (2015). The sense of confidence during probabilistic learning: A normative account. *PLoS Computational Biology*, *11*(6), e1004305.
Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, *15*(7), 1040–1046.
Nelson. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109.

Nelson, B., Whitford, T. J., Lavoie, S., & Sass, L. A. (2014). What are the neurocognitive correlates of basic self-disturbance in schizophrenia?: Integrating phenomenology and neurocognition. Part 1 (Source monitoring deficits). *Schizophrenia Research*, *152*(1), 12–19.

Nielsen, T. (1963). Volition: A new experimental approach. *Scandinavian Journal of Psychology*, *4*(4), 225–230.

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*(10), 1544–1553.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157.

Nobre, A. C. (Kia), & Kastner, S. (Eds.). (2014). *The Oxford Handbook of Attention*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199675111.001.0001

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*(2), 160.

Preuschoff, K., 't Hart, B. M., & Einhäuser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience*, *5*, 115.

Rahnev, D., Desender, K., Lee, A. L., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L. Y., Balcı, F., & Bang, J. W. (2020). The confidence database. *Nature Human Behaviour*, *4*(3), 317–325.

Rausch, M., & Zehetleitner, M. (2019). The folded X-pattern is not necessarily a statistical signature of decision confidence. *PLoS Computational Biology*, *15*(10), e1007456.

Rescorla, R. A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical Conditioning, Current Research and Theory*, *2*, 64–69.

Rizzolatti, G., Riggio, L., & Sheliga, B. M. (1994). Space and selective attention. *Attention and Performance XV*, *15*, 231–265.

Rolfs, M., & Schweitzer, R. (2022). Coupling perception to action through incidental sensory consequences of motor behaviour. *Nature Reviews Psychology*, *1*(2), 112–123.

Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human metacognition across domains: Insights from individual differences and neuroimaging. *Personality Neuroscience*, *1*, e17.

Salomon, R., Kannape, O. A., Debarba, H. G., Kaliuzhna, M., Schneider, M., Faivre, N., Eliez, S., & Blanke, O. (2021). Agency Deficits in a Human Genetic Model of Schizophrenia: Insights From 22q11DS Patients. *Schizophrenia Bulletin*, sbab143. https://doi.org/10.1093/schbul/sbab143

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron*, *90*(3), 499–506.

Schmack, K., Bosc, M., Ott, T., Sturgill, J. F., & Kepecs, A. (2021). Striatal dopamine mediates hallucination-like perception in mice. *Science*, *372*(6537), eabf4740. https://doi.org/10.1126/science.abf4740

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Shadmehr, R., Reppert, T. R., Summerside, E. M., Yoon, T., & Ahmed, A. A. (2019). Movement vigor as a reflection of subjective economic utility. *Trends in Neurosciences*, *42*(5), 323–336.

Shahar, N., Moran, R., Hauser, T. U., Kievit, R. A., McNamee, D., Moutoussis, M., NSPN Consortium, & Dolan, R. J. (2019). Credit assignment to state-independent task

representations and its relationship with model-based decision making. *Proceedings of the National Academy of Sciences*, *116*(32), 15871–15876.
Shekhar, M., & Rahnev, D. (2021). Sources of metacognitive inefficiency. *Trends in Cognitive Sciences*, *25*(1), 12–23.
Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177.
Stern, Y., Koren, D., Moebus, R., Panishev, G., & Salomon, R. (2020). Assessing the Relationship between Sense of Agency, the Bodily-Self and Stress: Four Virtual-Reality Experiments in Healthy Individuals. *Journal of Clinical Medicine*, *9*(9), 2931.
Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition*, *17*(1), 219–239.
Thomas, A. W., Molter, F., Krajbich, I., Heekeren, H. R., & Mohr, P. N. (2019). Gaze bias differences capture individual choice behaviour. *Nature Human Behaviour*, *3*(6), 625–635.
Ting, C.-C., Salem-Garcia, N., Palminteri, S., Engelmann, J. B., & Lebreton, M. (2023). Neural and computational underpinnings of biased confidence in human reinforcement learning. *Nature Communications*, *14*(1), 6896.
Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766.
Turk-Browne, N. B., Yi, D.-J., & Chun, M. M. (2006). Linking implicit and explicit memory: Common encoding factors and shared representations. *Neuron*, *49*(6), 917–927.
Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications*, *8*(1), 14637.
Van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). Confidence is the bridge between multi-stage decisions. *Current Biology*, *26*(23), 3157–3168.
van Ede, F., Board, A. G., & Nobre, A. C. (2020). Goal-directed and stimulus-driven selection of internal representations. *Proceedings of the National Academy of Sciences*, *117*(39), 24590–24598.
Van Ede, F., Chekroud, S. R., & Nobre, A. C. (2019). Human gaze tracks attentional focusing in memorized visual space. *Nature Human Behaviour*, *3*(5), 462–470.
Van Ede, F., Chekroud, S. R., Stokes, M. G., & Nobre, A. C. (2019). Concurrent visual and motor selection during visual working memory guided action. *Nature Neuroscience*, *22*(3), 477–483.
Viganò, S., Bayramova, R., Doeller, C. F., & Bottini, R. (2024). Spontaneous eye movements reflect the representational geometries of conceptual spaces. *Proceedings of the National Academy of Sciences*, *121*(17), e2403858121.
Vossel, S., Mathys, C., Daunizeau, J., Bauer, M., Driver, J., Friston, K. J., & Stephan, K. E. (2014). Spatial attention, precision, and Bayesian inference: A study of saccadic response speed. *Cerebral Cortex*, *24*(6), 1436–1450.
Weidemann, C. T., & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. *Royal Society Open Science*, *3*(4), 150670.
Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547. https://doi.org/10.7554/eLife.49547
Wolfe, J. M. (2020). Visual search: How do we find what we are looking for? *Annual Review of Vision Science*, *6*, 539–562.