# Title Page

**Title:** False alarms are driven by dynamic predictive templates

**Authors**:

Veith Weilnhammer, Yuki Murai, David Whitney

**Affiliations and Contribution**:

Helen Wills Neuroscience Institute, University of California Berkeley, USA  
 Center for Information and Neural Networks, Osaka, Japan

**Corresponding Author**:

Veith Weilnhammer, Helen Wills Neuroscience Institute, University of California Berkeley, USA, email: [veith.weilnhammer@gmail.com](mailto:veith.weilnhammer@gmail.com)  
  
 **Word Count**: 885 words (main text without methods)

# Main

False alarms occur when participants report having perceived a stimulus that was, in fact, never presented. False alarms share this feature with hallucinations—vivid and transient experiences of objects, such as images or sounds, that occur in the absence of a cause in the external world—and have therefore developed into one of the most influential paradigms to study perceptual symptoms in psychosis1–3. Yet, to serve as a proxy for hallucinations, false alarms need to satisfy at least two more criteria: First, false alarms should not only represent the erroneous report of a signal, but reflect perceptual experiences that are characterized by specific content2. Second, false alarms should occur at a timescale compatible with the temporal dynamics of hallucinations, which often unfold on the order of seconds to minutes, particularly in early stages of psychotic illness4. In this work, we validate both criteria by combining Hidden Markov Models5 (HMMs) with a classification image approach6 to show that false alarms are more likely to occur in an internal mode of perception, a minute-long state of the brain, during which the content of perception is strongly biased toward previous experiences.

According to Bayes theorem, false alarms () become likely when people expect to see a signal (, prior), and when the features of a noisy input are compatible with the expected signal (, likelihood). In natural environments, where the recent past is a predictor of the near future, previous stimuli may induce expectations about what is likely to be perceived next, and generate predictive templates that distort the perception of compatible noise into the experience of a signal ().

To test the contribution of prior and likelihood to false alarms, we analyzed data from 22 participants who judged whether close-to vertical gratings (targets) were present (alarms) or absent (rejections) in white noise images. The alarm rate depended on three factors: the contrast of the target at the current trial ( ± , z = , p = ), the contrast of the target at the preceding trial ( ± , z = , p = ), and the relative power of the current noise image at close-to-vertical orientations ( ± , z = , p = ).

This confirms that false alarms are driven by predictive templates, i.e, they were more likely to occur when previous stimuli induced a strong signal expectation (, prior), and when the sensory noise at the current trial matched the features of the expected signal (, likelihood). To visualize these predictive templates, we subtracted the noise power at rejection trials from the noise power at alarm trials. The resulting classification images6 revealed a noise power peak at close-to-vertical-orientations (Figure X, ± , T() = , p = ) that matched the average orientation of the target gratings. These findings confirm criterion 1: at the time of false alarms, observers have perceptual experiences of a specific content - in this case, gratings with a close-to-vertical orientation. The hallucinated content is determined by predictive templates that are induced by preceding experiences.

If predictive templates reflect the mechanism of hallucinations, then their strength should fluctuate at a timescale compatible with the duration of hallucinatory experiences. To assess the temporal dynamics of the predictive templates, we estimated HMMs that inferred transitions between two latent states, each state linked to a logistic regression model that predicted trial-wise responses (alarms versus rejections) from the target contrast at the current trial () and the response at the preceding trial (, please note that these logistic regression models were independent of the presented noise, and target orientation). In line with previous results5,7, the HMMs revealed slow alternations between an external mode, where perception closely followed external stimulus (), and an internal mode, where perception was strongly biased by preceding experiences ( > , relative to a one-state control model: -1156). Participants spent ± % of their time in internal mode, with alternation between modes occurring in intervals of ± trials, corresponding to ± sec.

The internal mode increased the rate of false alarms relative to external mode ( ± , T() = , p = , Figure X). In the internal mode, perception depended more on the stimuli presented at the preceding trial ( ± , z = , p = ) and less on the stimulus presented at the current trial ( ± , z = , p = ).  
Alarms during the internal mode were associated with a close-to-vertical noise-power-peak that was shifted toward the orientation of the stimulus presented at the preceding trial ( ± , T() = , p = ). This shift was not observed during the external mode ( ± , T() = , p = , Figure X). These findings confirmed criterion 2: the predictive templates that drive false alarms are dynamic, carry content, and fluctuate at a timescale compatible with the duration of individual hallucinatory experiences.

By combining HMMs with the classification image approach5,6, we show that the false alarms observed in our study are perceptual phenomena, characterized by specific contents, and driven by dynamic predictive templates, governed by transitions between external and internal modes of perception that alternate on a timescale compatible with the on- and offset of hallucinatory experiences. Future research should explore whether false alarms, and ultimately clinical hallucinations, can be mitigated by interventions that disrupt the internal mode.

# Methods

## Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Veith Weilnhammer ([veith-andreas.weilnhammer@charite.de](mailto:veith-andreas.weilnhammer@charite.de)).

## Materials availability

This study did not generate new unique reagents.

## Data and code availability

All custom code and behavioral data are available on <https://github.com/veithweilnhammer/Modes>. This manuscript was created using the *R Markdown* framework, which integrates all data-related computations and the formatted text within one document. With this, we wish to make our approach fully transparent and reproducible for reviewers and future readers.

## Experimental model and subject details

#### Apparatus

The participants viewed visual stimuli on a gamma-corrected CRT monitor (Sony Trinitron Multiscan G520, 1024x768 pixels, 100 Hz refresh rate, Konica Minolta LS-110 luminometer for gamma correction) at a viewing distance of 57.3 cm, with head stabilization provided by a chin rest. G

### Stimuli and Procedure

All stimuli were presented using MATLAB (MathWorks, R2017a) and the Psychophysics Toolbox. In 44% of the trials, we presented low-contrast near-to-vertical Gabor patches which were embedded in static white noise. In another 44% of the trials, we presented only static white noise. The remaining 12% of the trials featured a high-contrast Gabor patch (40% Michelson contrast) embedded in white noise. This high-contrast inducer was introduce to boost the phenomenon of serial dependency. The inducers were oriented 10 degrees clockwise or counterclockwise relative to vertical, and presented in intervals randomized between 4 and 10 trials. Stimuli were displayed for 500 ms, with a maximum white noise contrast of 60%. To reduce potential luminance aftereffects, the target was masked by low-pass filtered luminance contrast noise (100% contrast) for 500 ms, and the Gabor phase was randomized across trials. Visual stimuli measured 14 x 14 degrees of visual angle (d.v.a.), with Gabor patches confined to a Gaussian contrast envelope with a 3 d.v.a. standard deviation. The spatial frequency of the Gabor patches was 0.5 cycles per degree (cpd), with orientation randomized within ±3 degrees relative to vertical. Participants judged the presence or absence of the near-vertical target Gabor in the noisy image. We adjusted the target contrast on a participant-by-participant basis to achieve a d’ of approximately 1.5 using data from preliminary experiments. Each session consisted of 100 trials, with inter-trial intervals randomized between 800-1200 ms. Participants completed 15-25 sessions in about 2 hours.

### Quantification and Statistical Analysis

Responses were categorized into one of four stimulus-response types: Hits (alarms on stimulus trials), correct rejections (rejections on no-stimulus trials), false alarms (alarms on no-stimulus trials), and misses (rejections on stimulus trials). We employed standard logistic and linear regression using R-packages lmer, glmer, and afex (see Supplemental Table S1).

#### Classification images

Following previous approaches6, we filtered each static white noise image with the Gaussian stimulus envelope, applied fast 2-D Fourier-transformation, and extracted the power at the Gabor patch’s spatial frequency, separately for orientations from 0 to 180 degress in three degree intervals. For every image, we computed the average power between 60 to 120 degrees, relative to the overall power from compelte orientation range. To visualize the classification images, we subtracted the average power-by-orientation at rejection trials from the average power-by-orientation at alarm trials. To examine serial dependence, the same analysis was conducted separately for trials following clockwise (CW) and counter-clockwise (CCW) inducers. In each condition and participant, we determined the average vector of the clasification image by computing the centroid of a polygon composed of vector endpoints in polar coordinates.

#### Hidden Markov Modeling

We used General Linear Models (GLM) to predict the response (0: rejection: 1: alarm) from the contrast at the current trial (: zero contrast, low contrast, high contrast) and the response at the preceding trial :

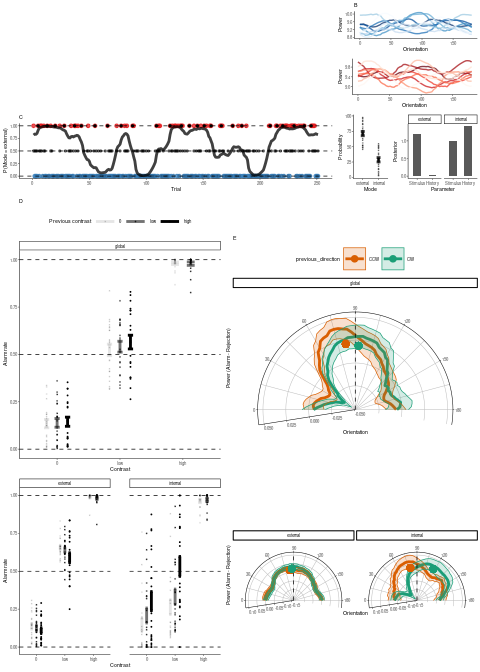
We then used Hidden Markov Models (HMMs, as implemented in the *SSM: Bayesian learning and inference for state space models* python library, <https://github.com/lindermanlab/ssm>) to model transitions between two latent states and . At each trial, the HMM predicted the latent state was or , with alternation between and that were governed by a 2 x 2 transition matrix :

Each latent state was linked to an independent GLM that predicted from and based on the weights :

For both states, parameters were initialized at the regression weights of a one-state control GLM. To compare the performance of the two-state GLM relative to the one-state control GLM, we computed the difference in Bayesian Information Criterion (BIC) between the two models. The SSM hyperparameters were defined as follows: = 10 (variance over the GLM weights , and = 1 (Dirichlet prior over the transition matrix ) was set to 1. The latent states and were linked to mode by comparing and (external mode: > , internal mode: > ). We labelled trial as external when .

# Figures

## Figure 1



**Figure 1.**

**A. Paradigm.** Participants were instructed to report whether they perceived (close-to-) vertical gratings embedded in noise. At stimulus trials, gratings were presented either at low or at high contrast. At no-stimulus trials, no grating was presented (contrast = 0). We hypothesized that alarms would be more likely when (i), targets are presented at high contrast, when (ii), previous stimuli are strong enough to induce perceptual expectations that bias the interpretation of the noisy input via the phenomenon of serial dependency, and when (iii), the presented noise is compatible with the average features of the target Gabor.

**B. Predictive perceptual templates.** We reasoned that alarms would be more frequent when participant expect to experience a signal, and when the sensory noise is compatible with the predicted signal ((, upper panel). We induced perceptual expectations () via the phenomenon of serial dependency (lower panel). Serial dependencies cause perception to be biased toward preceding experiences, causing alarms to be more likely to be followed by alarm, and rejections more likely to be followed by rejection. Since serial dependencies are known to be stronger after confident experiences, we predicted that false alarms would be more frequent after high-contrast stimuli. To compute the compatibility of the white noise with the target gabor (), we subtracted the power-by-orientation distribution of the random white noise images at rejection trials (blue) from the power-by-orientation distribution of the random noise images at alarm trials (red, right panel).

**C. External and internal modes in perception.** Perception is known to slowly alternate between two opposing modes (left panel, example data from one representative participant): During external mode, perception (red) is dominated by incoming stimuli (blue). By contrast, during internal mode, perception is shaped by the perceptual expectations induced by preceding experiences. Between-mode alternations (black line indicated the probability of external mode) can by discovered by Hidden Markov Models (HMM). HMMs describe alternations between two independent General Linear Models (GLMs), each of which predicts perceptual decisions via weights assigned to the stimulus presented at the current trial, and experience made at the preceding trial. In this experiment, the GLM-HMM discovers an external mode, during which the weight assigned to incoming stimuli is high, while the weight assigned to the history of previous responses is low (middle and right panels).

**E. The effect of predictive templates on alarms and rejections depend on mode.** Globally, alarm rate correlated positively with the contrast of stimuli presented at the preceding trial ( ± , z = , p = ). This suggests that serial dependencies contribute to false alarms. When separating our data in episodes of external and internal modes, we found positive serial dependencies only during the internal mode ( ± , z = , p = ). During the external mode, by contrast, we observed a negative correlation between alarm rate and preceding stimulus contrast ( ± , z = , p = ).

**E. The effect of predictive templates on alarms and rejections depends on mode.** Globally, the alarm classification images revealed a peak around 90°, mirroring the orientations of the average target gabor ( ± , T() = , p = ). During the internal mode, we found a shift of the close-to-vertical noise power peak toward the orientation of the preceding stimulus ( ± , T() = , p = ; CW in green, CCW in orange). This was not observed during the external mode ( ± , T() = , p = ).

# Supplementary Information

## Supplemental Table S1

| RESOURCE | SOURCE | IDENTIFIER |
| --- | --- | --- |
| **Deposited data & code** |  |  |
| Analyzed data & custom code | <https://github.com/veithweilnhammer/XYZ/> | N/A |
| **Software** |  |  |
| **Python 3** | <http://www.python.org/> | RRID:SCR\_008394 |
| Jupyter Notebook | <https://jupyter.org/> | RRID:SCR\_018315 |
| numpy | [http://www.numpy.org](http://www.numpy.org/) | RRID:SCR\_008633 |
| pandas | [https://pandas.pydata.org](https://pandas.pydata.org/) | RRID:SCR\_018214 |
| SSM | <https://github.com/lindermanlab/ssm> | N/A |
| **Matlab** | <https://www.mathworks.com/> | RRID:SCR\_001622 |
| Psychtoolbox 3 | <http://psychtoolbox.org/> | RRID:SCR\_002881 |
| **R** | <http://www.r-project.org/> | RRID:SCR\_001905 |
| RStudio | <https://www.rstudio.com/> | RRID:SCR\_000432 |
| lme4, afex, ggplot2, ggridges, gridExtra, tidyr, plyr | <http://cran.r-project.org/> | RRID:SCR\_003005 |

# References

1. Powers, A. R. *et al.* [Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors](https://doi.org/10.1126/science.aan3458). *Science* **357**, 596–600 (2017).

2. Horga, G. *et al.* [An integrative framework for perceptual disturbances in psychosis](https://doi.org/10.1038/s41583-019-0234-1). *Nature Reviews Neuroscience* **20**, 763–778 (2019).

3. Schmack, K. *et al.* [Striatal dopamine mediates hallucination-like perception in mice](https://doi.org/10.1126/science.abf4740). *Science (New York, N.Y.)* **372**, eabf4740 (2021).

4. Hermans, K. *et al.* [Temporal dynamics of suspiciousness and hallucinations in clinical high risk and first episode psychosis](https://doi.org/10.1016/j.psychres.2020.113039). *Psychiatry Research* **290**, 113039 (2020).

5. Weilnhammer, V. *et al.* N-Methyl-D-aspartate receptor hypofunction causes recurrent and transient failures of perceptual inference. (2024) doi:[10.1101/2024.05.24.595590](https://doi.org/10.1101/2024.05.24.595590).

6. Murai, Y. *et al.* [Serial dependence revealed in history-dependent perceptual templates](https://doi.org/10.1016/j.cub.2021.05.006). *Current biology: CB* **31**, 3185–3191.e3 (2021).

7. Weilnhammer, V. *et al.* [Sensory processing in humans and mice fluctuates between external and internal modes](https://doi.org/10.1371/journal.pbio.3002410). *PLOS Biology* **21**, e3002410 (2023).