NEUROSCIENCE

# Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors

A. R. Powers,[1] C. Mathys,[2,3,4] P. R. Corlett[1]*

**Some people hear voices that others do not, but only some of those people seek treatment. Using a Pavlovian learning task, we induced conditioned hallucinations in four groups of people who differed orthogonally in their voice-hearing and treatment-seeking statuses. People who hear voices were significantly more susceptible to the effect. Using functional neuroimaging and computational modeling of perception, we identified processes that differentiated voice-hearers from non–voice-hearers and treatment-seekers from non–treatment-seekers and characterized a brain circuit that mediated the conditioned hallucinations. These data demonstrate the profound and sometimes pathological impact of top-down cognitive processes on perception and may represent an objective means to discern people with a need for treatment from those without.**

Perception is not simply the passive reception of inputs (*1*). We actively infer the causes of our sensations (*2*). These inferences are influenced by our prior experiences (*3*). Priors and inputs might be combined according to Bayes' rule (*4*). Prediction errors, the mismatch between priors and inputs, contribute to belief updating (*5*). Hallucinations (percepts without external stimulus) may arise when strong priors cause a percept in the absence of input (*6*). We tested this theory by engendering new priors about auditory stimuli in human observers using Pavlovian conditioning.

Even in healthy individuals, the repeated co-occurrence of visual and auditory stimuli can induce auditory hallucinations (*7*). We examined this effect with functional imaging. Some argue that in patients with psychosis, weak priors lead to aberrant prediction errors, resulting in auditory verbal hallucinations (AVH) (*8*). Others have observed strong priors in patients, but the effects were not specific to hallucinations (*9*, *10*). Such inconsistencies may reflect the hierarchical organization of perception: Perturbations may affect some levels of the hierarchy and not others (*9*). We used computational modeling to infer the strength of participants' hierarchical perceptual beliefs from their behavioral responses during conditioning (*11*). Our model captured how priors are combined with sensory evidence, allowing us to test the strong-prior hypothesis directly.

Participants worked to detect a 1-kHz tone occurring concurrently with presentation of a checkerboard visual stimulus. First, we determined individual thresholds for detection and psychometric curves (*12*). Then, at the start of conditioning, the tone was presented frequently at threshold (Fig. 1A, left), engendering a belief in audio-visual association. This belief was then tested (Fig. 1A, right) with increasingly frequent subthreshold and target-absent trials (Fig. 1B). Conditioned hallucinations occurred when subjects reported tones that were not presented, conditional upon the visual stimulus.

We recruited four groups of subjects (Fig. 1C): people with a diagnosed psychotic illness who heard voices (P+H+, $n$ = 15); those with a similar illness who did not hear voices (P+H–, $n$ = 14); an active control group who heard daily voices, but had no diagnosed illness (P–H+, $n$ = 15) (*13*)—they attributed their experiences metaphyscially (supplementary materials) (*14*); and last, controls without diagnosis or voices (P–H–, $n$ = 15).

Groups were matched demographically (tables S1 to S4). Rates of detection of tones at threshold

[1]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA. [2]International School for Advanced Studies (SISSA), Trieste, Italy. [3]Max Planck University College London (UCL) Centre for Computational Psychiatry and Ageing Research, London, UK. [4]Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland.
*Corresponding author. Email: philip.corlett@yale.edu

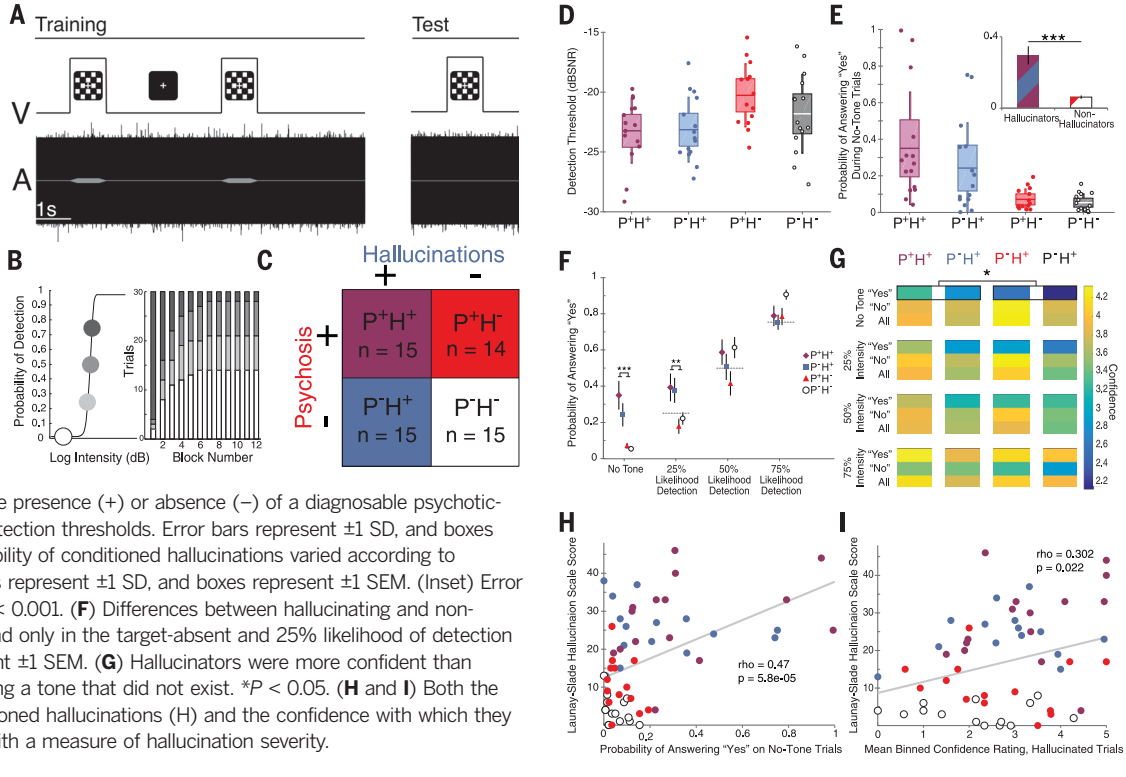


**Fig. 1. Methods and behavioral results.** (**A**) Trials consisted of simultaneous presentation of a 1000-Hz tone in white noise and a visual checkerboard. (**B**) We estimated individual psychometric curves for tone detection (left) and then systematically varied stimulus intensity over 12 blocks of 30 conditioning trials. Threshold tones were more likely early, and absent tones were more likely later (right). (**C**) Groups varied along two dimensions: the presence (+) or absence (–) of daily AVH (blue) and the presence (+) or absence (–) of a diagnosable psychotic-spectrum illness (red). (**D**) Detection thresholds. Error bars represent ±1 SD, and boxes represent ± 1 SEM. (**E**) Probability of conditioned hallucinations varied according to hallucination status. Error bars represent ±1 SD, and boxes represent ±1 SEM. (Inset) Error bars represent ±1 SEM. ***$P$ < 0.001. (**F**) Differences between hallucinating and non-hallucinating groups were found only in the target-absent and 25% likelihood of detection conditions. Error bars represent ±1 SEM. (**G**) Hallucinators were more confident than nonhallucinators when reporting a tone that did not exist. *$P$ < 0.05. (**H** and **I**) Both the probability of reporting conditioned hallucinations (H) and the confidence with which they were reported (I) correlated with a measure of hallucination severity.

were similar across groups. All groups demonstrated conditioned hallucinations. However, those with daily hallucinations endorsed more conditioned hallucinations than those without, regardless of diagnosis ($F_{1,55}$ = 19.59, $P$ = 5.82 × $10^{-5}$) (Fig. 1D). This effect remained after accounting for differences in detection thresholds (Fig. 1E, fig. S1, and table S5). Group differences in propensity to report tones were observed only in the "no-tone" and 25% "likelihood of detection" conditions (intensity-by-hallucination status $F_{3,165}$ = 13.59, $P$ = 5.73 × $10^{-4}$) (Fig. 1F).

Participants also rated their decision confidence by holding down the response button (Fig. 1G). Participant confidence varied with stimulus intensity ("yes": $R$ = 0.39, $P$ = 7.46 × $10^{-10}$; "no": $R$ = 0.22, $P$ = 9.02 × $10^{-4}$). However, hallucinators were more confident in their conditioned hallucinations than nonhallucinators ($F_{1,53}$ = 6.50, $P$ = 0.045). Both conditioned hallucinations and confidence correlated with hallucination severity outside of the laboratory (Fig. 1, H and I, and fig. S3).

In order to establish whether conditioned hallucinations involved true percepts, we first identified tone-responsive regions from thresholding runs [peaks at (−60, −20, 2) and (62, −28, 10)] (Fig. 2A). As observed with elementary hallucinations (*15*), activity in tone-responsive regions was greater during conditioned hallucinations compared with correct rejections ($t_{56}$ = 4.93, $P$ = 7.59 × $10^{-6}$) (Fig. 2B). Electrical stimulation of this region in human patients produces AVH (*16*). Taken together, these findings are consistent conditioned hallucinations involving actual perception.

Whole-brain analysis revealed that conditioned hallucinations also engaged anterior insula cortex (AIC), inferior frontal gyrus, head of caudate, anterior cingulate cortex (ACC), auditory cortex, and posterior superior temporal sulcus (STS) (Fig. 2C and table S6). A meta-analysis of symptom-capture–based studies examining neural activity of AVH highlighted similar regions (Fig. 2D) (*17*). AIC and

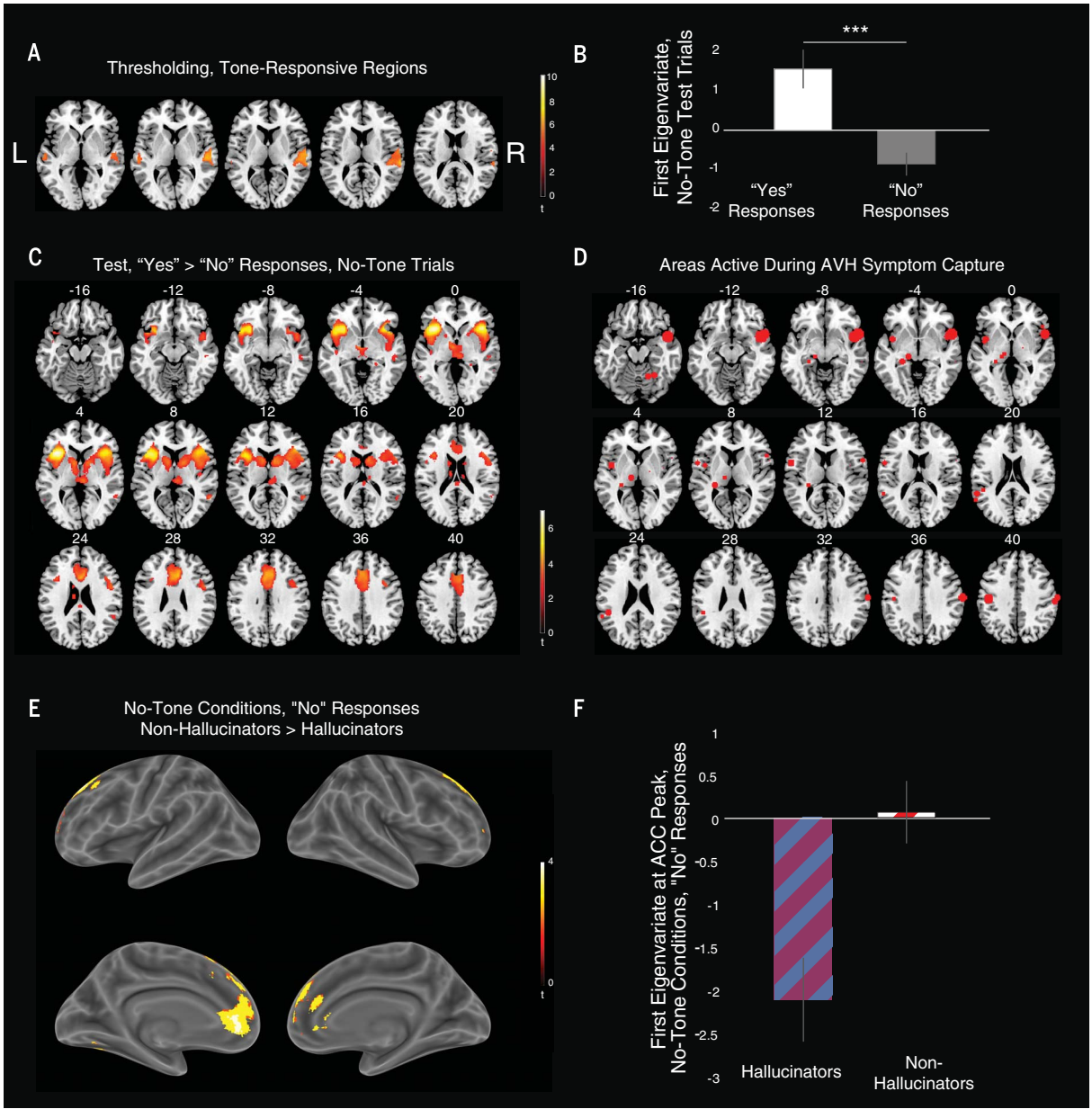


**Fig. 2. Imaging results.** (**A**) Bilateral supplemental auditory cortex covaried with tone intensity during thresholding (family-wise error rate–corrected, $P$ < 0.05). (**B**) Parameter estimates from this region showed increased activation during conditioned hallucinations. ***$P$ < 0.001. (**C**) Whole-brain analysis during conditioned hallucinations (false discovery rate–corrected, $P$ < 0.05). (**D**) Clusters derived from a meta-analysis (*17*) of AVH experiences during functional imaging. (**E** and **F**) Hallucinators were much less likely to engage ACC during correct rejections. Error bars represent ±1 SEM.

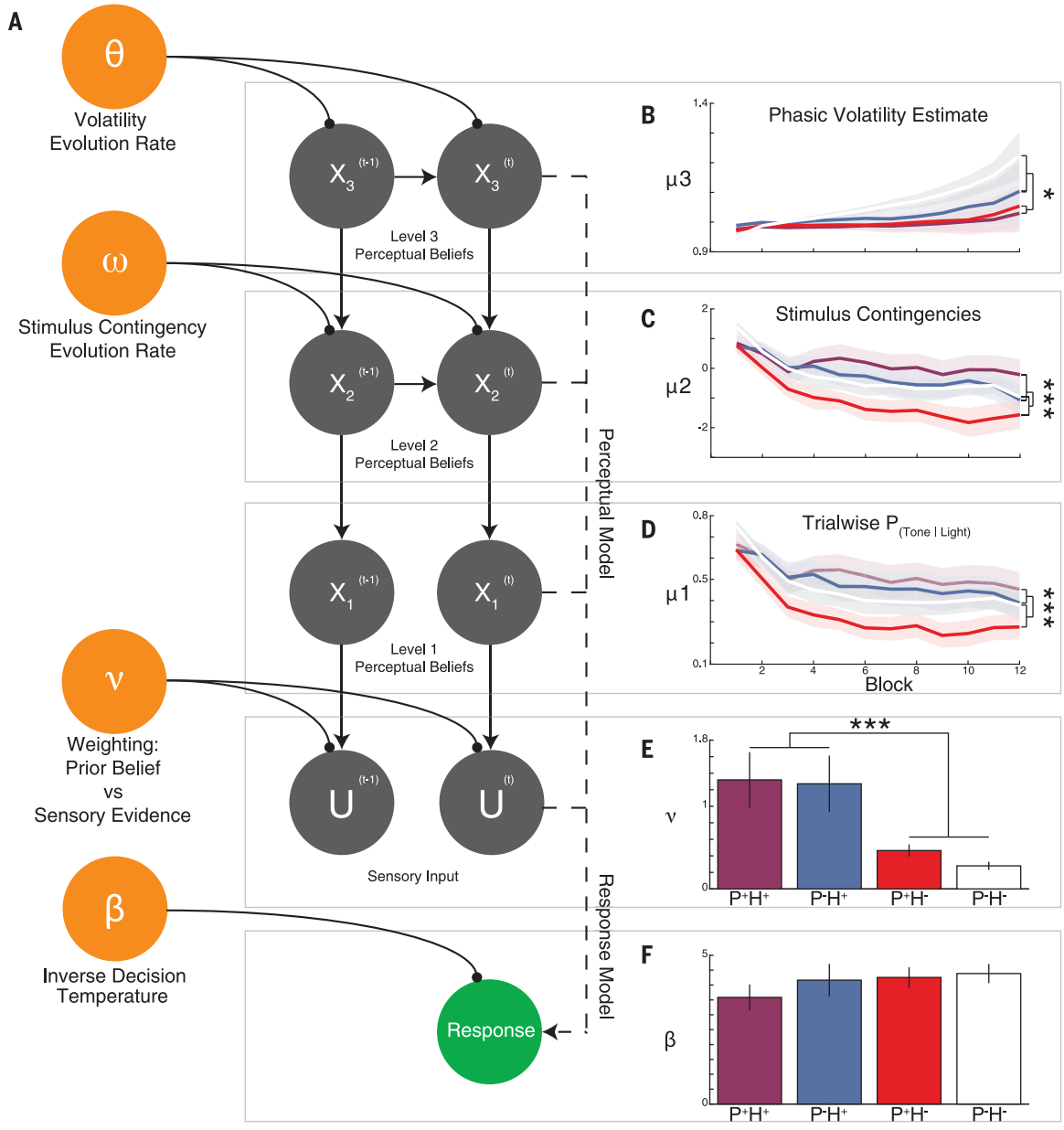


**Fig. 3. HGF analysis.** (**A**) Computational model, mapping from experimental stimuli to observed responses through perceptual and response models. The first level ($X_1$) represents whether the subject believes a tone was present or not on trial *t*. The second level ($X_2$) is their belief that visual cues are associated with tones. The third level ($X_3$) is their belief about the volatility of the second level. The HGF allows for individual variability in weighting between sensory evidence and perceptual beliefs (parameter ν). (**B**) At $X_3$, there was a significant block-by-psychosis interaction. *$P < 0.05$. (**C** and **D**) Significant block-by-hallucination status interactions were seen at layers (D) $X_1$ and (C) $X_2$. ***$P < 0.001$. (**E**) ν was significantly higher in those with hallucinations when compared with their nonhallucinating counterparts. ***$P < 0.001$. (**F**) No main effects of group or interaction effects were seen for the decision noise parameter within the response model. Error bars and line shadings represent ±1 SEM. Purple, P+H+; blue, P–H+; red, P+H–; white, P–H–.

ACC responses frequently correlate with stimulus salience (*18*). However, their activation before near-threshold stimulus presentation predicts detection (*19*). Caudate is engaged during audiovisual associative learning (*20*). Likewise, AIC and ACC are engaged during multisensory integration (*21*).

There were no significant between-group differences in brain responses during conditioned hallucinations. However, hallucinators deactivated ACC more [peak at (–16, 54, 14); cluster-extent thresholded, starting value 0.005, critical cluster extent ($k_e$) = 99] during correct rejections compared with nonhallucinators (Fig. 2, E and F).

To further dissect conditioned hallucinations, we modeled their underlying computational mechanisms (Fig. 3A) using the hierarchical Gaussian filter (HGF) (*11*). We defined a perceptual model consisting of low-level perceptual beliefs ($X_1$), visual-auditory associations ($X_2$), and the volatility of those associations ($X_3$), as well as evolution rates encoding the relationships between levels (ω, θ). Critically, our perceptual model allowed for variability in weighting between sensory evidence and perceptual beliefs (ν). For ν = 1, prior and observation have equal weight; for ν > 1, the prior has more weight than that of the observation (strong priors); and for ν < 1, the observation has more weight than that of the prior (weak priors). The resultant posterior probability of a tone is then fed to a separate response model.

Model parameters were fit to behavioral data, and the model was optimized by using log model evidence and simulations of observed behavior (figs. S3 and S4). Mean trajectories of perceptual beliefs were compared across groups (Fig. 3, B to D). Participants with hallucinations exhibited stronger beliefs at levels 1 ($X_1$: $F_{11,605} = 4.8$, $P = 3.89 \times 10^{-7}$) (Fig. 3D) and 2 ($X_2$: $F_{11,605} = 3.89$, $P = 1.84 \times 10^{-5}$) (Fig. 3C). $X_3$ beliefs evolved less in those with psychosis, who failed to recognize the increasing volatility in contingencies ($F_{11,605} = 2.11$, $P = 0.018$) (Fig. 3A).

Consistent with strong-prior theory, ν was significantly larger in those with hallucinations when compared with their nonhallucinating counterparts (Fig. 3E), regardless of diagnosis ($F_{1,55} = 13.96$, $P = 4.45 \times 10^{-4}$). Response model parameters did not differ across the groups (Fig. 3F).

We regressed model parameters onto task-induced brain responses (Fig. 4A). The $X_1$ trajectory

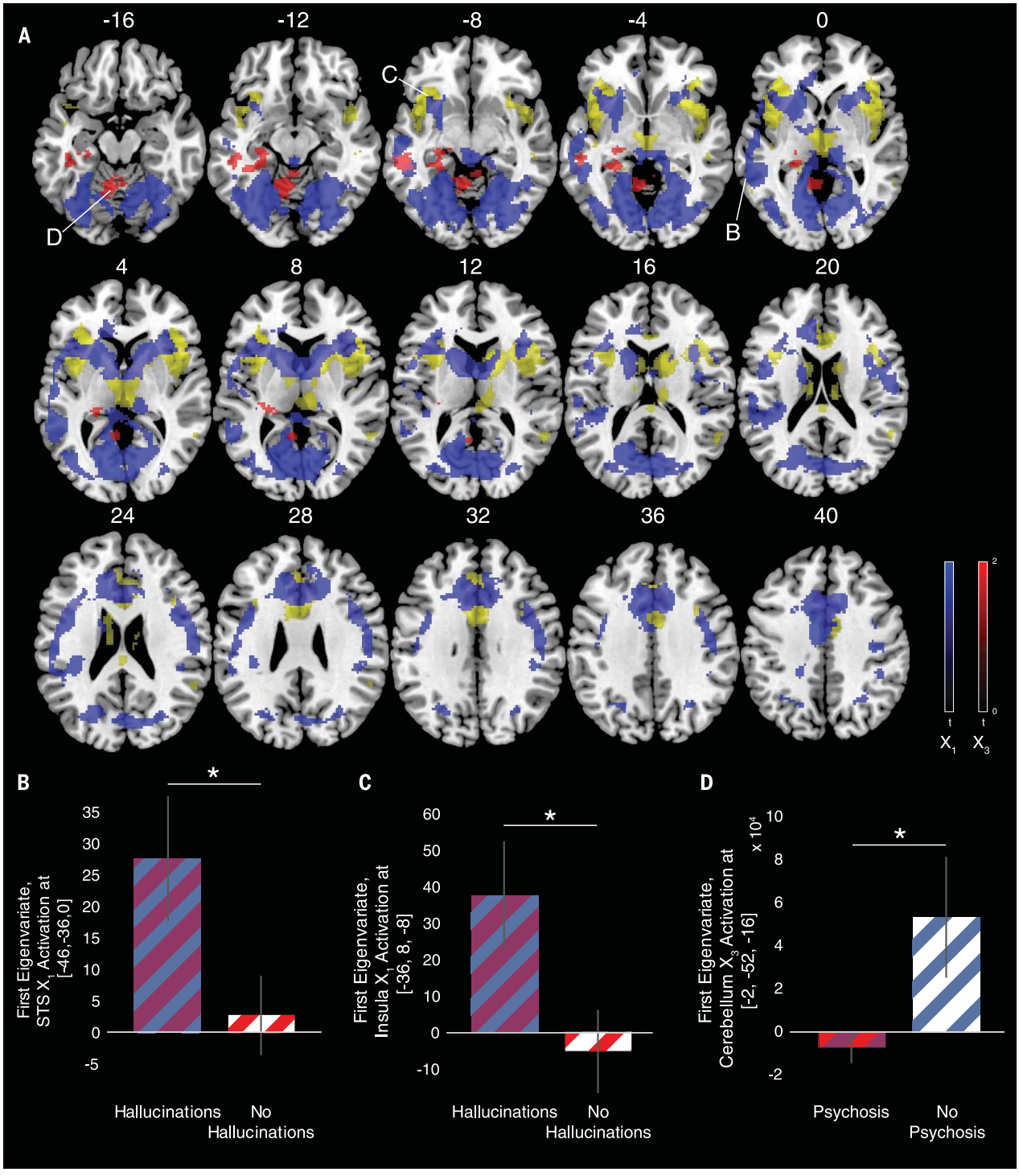


**Fig. 4. HGF imaging results.** (**A**) HGF trajectories for $X_1$ (blue) and $X_3$ (red) regressed onto blood oxygen level–dependent time courses for the conditioned hallucinations task. Regions that identified significantly active during conditioned hallucinations (from Fig. 3C) are highlighted in yellow for reference. All images are cluster-extent thresholded at starting value 0.05; critical $k_e$ for $X_1$ = 545 and $X_3$ = 406. (**B** and **C**) Parameter estimates of $X_1$ fit extracted from 5-mm sphere centered on (B) STS and (C) anterior insula activation differ based on hallucination status. (**D**) Parameter estimates of $X_3$ fit extracted from 1-mm sphere centered on cerebellar vermis activation differ based on psychosis status. Error bars represent 1 SEM.

covaried with several conditioned hallucination-responsive regions, including STS (table S7). $X_3$ trajectories, by contrast, covaried with hippocampus/parahippocampal gyrus and medial cerebellum (table S8). Parameter estimates from the $X_1$-sensitive STS [(−46 −36, 0), $T_{57}$ = 2.09, $P$ = 0.042] (Fig. 4B) and AIC [(36, 8, −8), $T_{57}$ = 2.26, $P$ = 0.027] (Fig. 4C) were significantly greater in those with hallucinations versus those without. This is consistent with STS conferring auditory expectations that are responsive to incoming visual input (*22*). Parameter estimates from the $X_3$-responsive cerebellar vermis [(−2, −52, −16)] (Fig. 4D) were lower in participants with psychosis as compared with those without ($T_{57}$ = 2.05, $P$ = 0.045). In the model,

subjects with psychosis were significantly less sensitive to the changes in contingency as the task progressed. Psychotic symptoms are often associated with pathological rigidity. Belief-updating correlated with responses in the hippocampus and cerebellum. Hippocampal activity correlates with uncertainty in perceptual predictions (*23*). The cerebellum has likewise been associated with production and updating of predictive models (*24*).

Our $X_1$, $X_2$, and ν findings are consistent with a strong-prior theory of hallucinations. The $X_3$ findings in psychotic patients may reflect a strong prior that contingencies are fixed. On the other hand, they could reflect a weak prior on volatility. These beliefs were not associated with hallucinations but rather psychosis more broadly. Under chronic uncertainty, secondary to consistent belief violation, it may be adaptive to resist updating beliefs (*25*).

Consistent with previous work applying signal detection theory (SDT) to AVH (*26*), we found liberal criteria and low perceptual sensitivity in our H+ groups. A liberal criterion may reflect poor reality monitoring (*26*). However, meta-d' (a metric of participants' meta-cognitive sensitivity) did not differ significantly between groups (fig. S6). SDT is a descriptive tool that does not distinguish aberrant perceptions from decisions. Our modeling work, however, localized group differences to the perceptual model alone. The prior weighting parameter (ν) distinguished H+ from H− groups and also predicted confidence in conditioned hallucinations (fig. S7). Our observations support an explanation of hallucinations based on strong perceptual priors. They suggest precision treatments for hallucinations, such as targeting cholinergically mediated priors (*27*), and interventions to mollify psychosis more broadly, such as cerebellar transcranial magnetic stimulation (*28*).

## REFERENCES AND NOTES


1. H. v. Helmholz, *Treatise on Physiological Optics* (Voss, Hamburg, ed. 3, 1909).
2. K. Friston, *Philos. Trans. R. Soc. London B Biol. Sci.* **360**, 815–836 (2005).
3. R. P. Rao, D. H. Ballard, *Nat. Neurosci.* **2**, 79–87 (1999).
4. T. Bayes, *Biometrika* **45**, 296–315 (1958).
5. R. A. Adams, K. E. Stephan, H. R. Brown, C. D. Frith, K. J. Friston, *Front. Psychiatry* **4**, 47 (2013).
6. K. J. Friston, *Behav. Brain Sci.* **28**, 764–766 (2005).
7. D. G. Ellson, *J. Exp. Psychol.* **28**, 1–20 (1941).
8. G. Horga, K. C. Schatz, A. Abi-Dargham, B. S. Peterson, *J. Neurosci.* **34**, 8072–8082 (2014).
9. C. Teufel *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 13401–13406 (2015).
10. K. Schmack, M. Rothkirch, J. Priller, P. Sterzer, *Hum. Brain Mapp.* **38**, 1767–1779 (2017).
11. C. Mathys, J. Daunizeau, K. J. Friston, K. E. Stephan, *Front. Hum. Neurosci.* **5**, 39 (2011).
12. A. B. Watson, D. G. Pelli, *Percept. Psychophys.* **33**, 113–120 (1983).
13. H. Verdoux, J. van Os, *Schizophr. Res.* **54**, 59–65 (2002).
14. A. R. Powers 3rd, M. S. Kelley, P. R. Corlett, *Schizophr. Bull.* **43**, 84–98 (2017).
15. J. Pearson *et al.*, *eLife* **5**, e17072 (2016).
16. W. Penfield, P. Perot, *Brain* **86**, 595–696 (1963).
17. L. Zmigrod, J. R. Garrison, J. Carr, J. S. Simons, *Neurosci. Biobehav. Rev.* **69**, 113–123 (2016).
18. P. Sterzer, A. Kleinschmidt, *Brain Struct. Funct.* **214**, 611–622 (2010).
19. S. Sadaghiani, G. Hesselmann, A. Kleinschmidt, *J. Neurosci.* **29**, 13410–13417 (2009).
20. H. E. den Ouden, K. J. Friston, N. D. Daw, A. R. McIntosh, K. E. Stephan, *Cereb. Cortex* **19**, 1175–1185 (2009).
21. P. J. Laurienti *et al.*, *Hum. Brain Mapp.* **19**, 213–223 (2003).
22. A. R. Powers 3rd, M. A. Hevey, M. T. Wallace, *J. Neurosci.* **32**, 6263–6274 (2012).
23. A. M. Schiffer, C. Ahlheim, M. F. Wurm, R. I. Schubotz, *PLOS ONE* **7**, e36445 (2012).
24. S. S. Shergill *et al.*, *JAMA Psychiatry* **71**, 28–35 (2014).
25. M. P. Karlsson, D. G. Tervo, A. Y. Karpova, *Science* **338**, 135–139 (2012).
26. R. P. Bentall, P. D. Slade, *Br. J. Clin. Psychol.* **24**, 159–169 (1985).
27. D. M. Warburton, K. Wesnes, J. Edwards, D. Larrad, *Neuropsychobiology* **14**, 198–202 (1985).
28. K. L. Parker, N. S. Narayanan, N. C. Andreasen, *Front. Syst. Neurosci.* **8**, 163 (2014).


## ACKNOWLEDGMENTS

The authors dedicate this work to the memory and legacy of Ralph E. Hoffman, M.D. Additional thanks go to M. Kelley, A. Bianchi, S. Bhatt, and E. Feeney for technical assistance as well as A. Nidiffer, L. Marks, S. Woods, and J. Krystal for their advice. This work was supported by the Connecticut Mental Health Center (CMHC) and Connecticut State Department of Mental Health and Addiction Services (DMHAS). P.R.C. was funded by an International Mental Health Research Organization/Janssen Rising Star Translational Research Award; National Institute of Mental Health (NIMH) grant 5R01MH067073-09; Clinical and Translational Science Award grant UL1 TR000142 from the National Center for Research Resources (NCRR) and the National Center for Advancing Translational Science (NCATS), components of the National Institutes of Health (NIH); NIH roadmap for Medical Research; the Clinical Neurosciences Division of the U.S. Department of Veterans Affairs; and the National Center for Post-Traumatic Stress Disorders, VA Connecticut Healthcare System (VACHS), West Haven, CT, USA. The contents of this work are solely the responsibility of the authors and do not necessarily represent the official view of NIH or the CMHC/DMHAS. A.R.P. was supported by the Integrated Mentored Patient-Oriented Research Training (IMPORT) in Psychiatry grant (5R25MH071584-07) as well as the Clinical Neuroscience Research Training in Psychiatry grant (5T32MH19961-14) from NIMH and a VA Schizophrenia Research Special Fellowship from VACHS, West Haven, CT, USA. Additional support was provided by the Yale Detre Fellowship for Translational Neuroscience as well as the Brain and Behavior Research Foundation in the form of a National Alliance for Research on Schizophrenia and Depression Young Investigator Award for A.R.P. The authors declare no conflicts of interest. Model code and data are stored at ModelDB (http://modeldb.yale.edu/229278). Imaging data are stored at NeuroVault (http://neurovault.org/collections/OCFEJCQE/).

## SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/357/6351/596/suppl/DC1
Materials and Methods
Supplemental Text
Figs. S1 to S7
Tables S1 to S8
References (*29*, *30*)

31 March 2017; accepted 4 July 2017
10.1126/science.aan3458