Title: Receptors, circuits and neural dynamics for prediction

Authors: Sounak Mohanta[1], Mohsen Afrasiabi[1], Cameron Casey[2], Sean Tanabe[2], Michelle J. Redinbaugh[1], Niranjan A. Kambi[1], Jessica M. Phillips[1], Daniel Polyakov[2], William Filbey[2], Joseph L. Austerweil[1], Robert D. Sanders[2*]†, Yuri B. Saalmann[1*]†

Affiliations:

[1] Department of Psychology, University of Wisconsin - Madison

[2] Department of Anesthesiology, University of Wisconsin - Madison

*Correspondence to: saalmann@wisc.edu or robert.sanders@wisc.edu

† Authors contributed equally

Abstract: Learned associations between stimuli allow us to model the world and make predictions, crucial for efficient behavior; e.g., hearing a siren, we expect to see an ambulance and quickly make way. While theoretical and computational frameworks for prediction exist, circuit and receptor-level mechanisms are unclear. Using high-density EEG and Bayesian modeling, we show that trial history and frontal alpha activity account for reaction times (a proxy for predictions) on a trial-by-trial basis in an audio-visual prediction task. Low-dose ketamine, a NMDA receptor blocker – but not the control drug dexmedetomidine – perturbed predictions, their representation in frontal cortex, and feedback to posterior cortex. This study suggests predictions depend on frontal alpha activity and NMDA receptors, and ketamine blocks access to learned predictive information.

One Sentence Summary: Predictions depend on NMDA receptors, representation in frontal cortex, and feedback to sensory cortex for comparison with sensory evidence.

Main Text: The classical view of sensory processing focuses on feedforward information transmission from the sensory organs to higher-order cortex, to generate representations of the world (*1*, *2*). However, expectations can strongly influence perception and behavior (*3*, *4*). This is captured in a radically different view of sensory processing, called predictive coding (PC), where the brain uses generative models to make inferences about the world (*5–9*), possibly even to support conscious experience (*3*, *4*, *10*). PC proposes that models represented in higher-order cortex transmit predictions to lower-order cortex along feedback connections. Any mismatch between feedback predictions and feedforward observed sensory evidence generates an error signal, leading to model updating (*11–13*). N-methyl-D-aspartate receptors (NMDARs) may play a key role in PC as they modulate higher-order (frontal) cortical excitability (*14–17*) and are enriched on the postsynaptic terminals of feedback connections (*18*). Ketamine, a NMDAR blocker (*19*), can reduce error signals, measured as auditory mismatch negativity (MMN) (*20–22*).

However, the NMDAR contribution to predictions and their neural representation are unclear.

To test circuit and receptor-level mechanisms of PC, we recorded 256-channel EEG of subjects performing an audio-visual delayed match-to-sample task (Fig. 1A). The task design separates predictions (generated during the delay period) from error processing (after image onset), which are not readily separable in oddball paradigms. Subjects performed the task before, during, and after recovery from sub-hypnotic dosing of ketamine, targeted to concentrations that modulate NMDARs, or the control drug dexmedetomidine (DEX), an α_2 adrenergic receptor agonist, selected to account for changes in arousal and modulation of hyperpolarization-activated cyclic nucleotide channels (HCN-1, which mediate ketamine's anesthetic effects (*23*)). Subjects initially learned paired associations (A1-V1, A2-V2, A3-V3) between 3 sounds (A1, A2, A3) and 3 images (V1, V2, V3) through trial-and-error, allowing us to modulate separate feedforward (auditory to frontal) and feedback (frontal to visual) pathways. During learning, each sound and image had equal probability (33%) of appearing in any given trial, preventing subjects developing any differential predictions due to stimulus frequency. Following presentation of both stimuli, subjects reported if the sound and image matched or not. To manipulate subjects' predictions during subsequent testing, we varied the probability of an image appearing after its associated sound. This probability was different for each sound: 85% chance of V1 after A1; 50% chance of V2 after A2; and 33% chance of V3 after A3. Thus, A1 was highly predictive (HP), A2 was moderately predictive (MP), and A3 was not match predictive (NP).

We hypothesized that increasing the predictive value of the sound would allow subjects to make better predictions about the upcoming image, enabling quicker responses (HP<MP<NP) in match trials. Reaction times (RTs) were faster when sounds had greater predictive value (ANOVA, N=11, P=0.001) (Fig. 1B). This result was further validated in parallel psychophysics experiments, where we controlled for possible match bias (*24*) (ANOVA, N=25, P<0.001; Fig. S1A). RTs could not be explained by speed-accuracy trade-offs, as subjects were most accurate for HP, followed by MP and NP sounds (ANOVA, N=25, P=0.001, Fig. S1B).

If predictions are mediated by NMDARs, then ketamine should block predictions; i.e., subjects administered a sub-anesthetic dose of ketamine should be unable to exploit the differential predictive value of each sound, causing the linear trend in RTs (HP<MP<NP sounds) to disappear. Subjects under ketamine (minimum effective plasma concentration of 0.2 μg/ml) no longer showed a linear correlation between the predictive value of sounds and RT (ANOVA, N=11, P=0.93) (Fig. 1C). This was not due to low accuracy as subjects' average accuracy was 79.93% under ketamine (87.19% without ketamine). Effects were specific to NMDAR manipulation, as DEX did not block predictions. Subjects under DEX showed significant (ANOVA, N=13, P=0.002) linear correlation between the predictive value of sounds and RT (Fig. 1E), similar to the baseline condition. Results were not due to sedation as, subjects were more alert under ketamine than dexmedetomidine

(ketamine average modified observer's assessment of alertness/sedation (OAA/S) score of 4.72 compared to 3.33 under DEX (5, awake – 1, unresponsive)). After recovery from ketamine (2-4 hours after ending ketamine administration, depending upon subject's recovery), subjects once again showed a linear correlation between the predictive value of sounds and RT (ANOVA, N=9, P=0.02; 2 subjects excluded post-ketamine due to vomiting; Fig. 1D). Overall, our results demonstrate predictions depend on NMDARs.

To identify what information in the trial history subjects base predictions on, we used a hierarchical drift-diffusion model (HDDM) (*25*), with two possible choices from our task (correct/incorrect). In the HDDM, evidence accumulates (drift process) from a starting point to two boundaries (one for each choice). It stops when it reaches a boundary, which determines the choice (and RT) for the trial (Fig. 1F). The starting point is determined by the predictive value of the sound, which may be biased towards one of the choices (reflected in a bias parameter; *z*). We used the causal power (*26*) of the sound-image association (calculated each trial based on prior trials) as a predictor of bias (Fig. 1G). (Causal power is the amount of evidence that a sound "causes" a particular image, as opposed to chance.) Specifically, we estimated the posterior probability density of the regression coefficient (β_1; Fig. 1F) to determine the relationship between bias and causal power. Bias was positively correlated with the predictive value of sounds (P=0.01; Fig. 1H). This suggests that more predictive sounds create a larger bias before the onset of the visual image and, as a result, decisions are reached more quickly, which translates to quicker RTs. Further, causal power (deviance information criterion, DIC=-3197) predicted RTs better than transitional probabilities (how often a particular image follows the sound only; DIC=-1795). This means subjects used the trial history beyond simple stimulus frequencies to generate predictions.

We used the HDDM to model drug effects. If ketamine blocks predictions, all sounds will generate similar biases; i.e., there will be no correlation between bias and the predictive value of sounds. Indeed, under ketamine, the posterior probability density of β_1 was not significantly different from zero (P=0.97; Fig. 1H). In contrast, under DEX, bias positively correlated with the predictive values of sounds (all sounds generating equal vs different biases, $P<10^{-10}$; Fig. 1H) similar to baseline. After recovery from ketamine, the posterior probability density of β_1 was significantly greater than zero ($P<10^{-10}$; Fig. 1H) confirming that subjects had again generated larger bias for more predictive sounds. The question is: (a) did subjects re-gain access to previously learned and stored predictive information (Fig. S2B) or (b) did they re-learn the predictive value of each sound after ketamine recovery (Fig. S2C)? To answer this, we used the HDDM to analyze the first 200 trials after recovery (to ensure each trial-type had at least 10 trials, for every subject). We found that only the former (option (a)) had significant (P=0.03; Fig. S2, D and E) positive posterior probability density. This suggests that ketamine did not erase previously learned predictive information, but rather ketamine prevented access to the predictive information.

We next investigated the circuit-level mechanism of prediction. Four clusters of electrodes, right frontal (RF), right central (RC), left central (LC) and occipital (OC),

showed significant modulation of delay period alpha power compared to baseline, irrespective of sound (Fig. 3A). Considering alpha power as an index of neural excitability (reduced alpha equating to reduced inhibition/increased excitability) (*27*, *28*), one might expect stronger predictions to be associated with lower alpha power, reflecting greater activation of prediction-encoding neurons. Indeed, stronger predictions significantly correlated with lower delay period alpha power at the RF electrode cluster (ANOVA, N=22, P=0.02; Fig. 2A-C; and Fig. S4A-C). This was not due to feedforward sensory processing, as all three sounds generated similar auditory ERPs (Fig. S3B). Hence, frontal alpha indexed predictions.

Because ketamine blocked predictions, subjects under ketamine should no longer show frontal alpha differences between sounds. NMDAR blockade has been shown to increase frontal cortical excitability (*14–17*), reducing response selectivity and signal-to-noise ratio (SNR) (*29*, *30*). We thus expect low-dose ketamine to increase frontal cortical excitability irrespective of the predictive value of sounds. This would manifest as similarly low RF alpha power for all sounds. Delay period alpha power at the RF electrode cluster was similar across sounds (ANOVA, N=10, P=0.87; Fig. 2, E-J). Whereas, under DEX, the RF alpha power still showed a significant linear trend (ANOVA, N=12, P=0.02; Fig. 2, I-K) similar to baseline. This suggests that NMDARs mediate prediction strength through modulation of frontal excitability (alpha).

Increased excitability of frontal cortex does not necessarily translate to useful prediction, if it reduces SNR. We trained a decoder to measure if this hyperexcitability, reflected in low RF alpha power, still allows differential representation of predictions. Before ketamine, the classification F score for each sound separated at 550 ms after sound onset (ANOVA, Holm's corrected P=0.0031; Fig. 3B). Similarly, classification F score for DEX separated at 600 ms after sound onset (ANOVA, Holm's corrected P=0.0017; Fig. 3D). In contrast, under ketamine, there was no separation of classification F score for each sound, and accuracy overall was lower (ANOVA, Holm's corrected P=0.31; Fig. 3C). The weighting of features contributing to classifier performance confirmed that, before ketamine, RF alpha power contributed most to classification accuracy (RF alpha power feature ($W_{RF\alpha}$) > other features ($W_{\sim RF\alpha}$), ANOVA, Holm's corrected P=0.020; Fig. 3F). This was also true under DEX ($W_{RF\alpha} > W_{\sim RF\alpha}$, ANOVA, P=0.027; Fig. 3H). As expected, under ketamine, RF alpha power contributed little to classification accuracy ($W_{RF\alpha} > W_{\sim RF\alpha}$, ANOVA, P=0.47; Fig. 3G). These results suggest that ketamine blocks predictions by reducing frontal SNR.

Although frontal activity correlates with predictions, we need to show that subjects use it. We used the HDDM to show RF alpha power predicts RT on a trial-by-trial basis. We calculated the posterior probability density of the regression coefficient, $\beta_{1\alpha}$ – the relationship between the bias and RF alpha power. Bias was inversely correlated with RF alpha power (P=0.02, Fig. 3E). This suggests that for more predictive sounds, lower RF alpha power creates a larger bias, and as a result, decisions are reached quicker (quicker RT). In contrast, ketamine blocked the correlation between bias and alpha power

(posterior probability distribution of regression coefficient, $\beta_{1\alpha}$, did not significantly differ from zero, P=0.69; Fig. 3H). This shows that subjects used frontal activity to make predictions.

The audio-visual task anatomically isolates predictions transmitted along feedback pathways to posterior visual areas, from the initial sound processing along ascending auditory pathways. We measured the predictive feedback from frontal to posterior cortex using spectral Granger causality. Granger causality increased after sound onset and remained elevated throughout the delay period. One might expect higher frontal excitability (from stronger predictions or effect of ketamine) to give rise to stronger feedback. Indeed, stronger predictions were proportionately associated with greater Granger causal influence of right frontal on right central electrodes during the delay period (ANOVA, N=22, P=0.006; Fig. 4, A-D). This suggests predictions are disseminated along feedback connections down the cortical hierarchy prior to image onset.

Because NMDAR blockers have been reported to perturb feedback pathways in macaques (*31*) and humans (*32*), we expected ketamine to alter feedback carrying predictions from frontal to posterior cortex. Under ketamine, there was no longer a correlation between the predictive value of sounds and alpha band Granger causal influence of right frontal on right central cortex (ANOVA, N=10, P=0.22; Fig. 4E); i.e., ketamine scrambled the feedback for each sound. In contrast, under DEX, the Granger causal influence in the alpha band still differed between sounds (ANOVA, N=12, P=0.02; Fig. 4F), similar to baseline. This suggests that the predictive feedback facilitating behavior depends on NMDARs.

In PC, divergence between predictive feedback and sensory evidence generates surprise conveyed in an error signal (when convergent, the prediction inhibits the sensory response). One might expect this to occur when the predicted and actual visual image differ. To test this, we compared the N170 peak of the V1 event-related potential for predictive (A1-V1) and not-predictive (A2-V1) trials in posterior cortex (T6 electrode). The MMN (A2-V1 minus A1-V1) during baseline is consistent with error signaling (paired T test, N=22, P=0.001; Fig. S5A). Blocking NMDAR-mediated prediction should reduce the MMN, as there will be no predictions in match (A1-V1) or non-match (A2-V1) trials. Indeed, the MMN disappeared under ketamine (P=0.18; Fig. S5B). Overall, this supports that predictions inhibit visual responses to expected images minimizing surprise.

Our results show NMDAR-mediated, circuit-level mechanisms of prediction. Frontal cortex represents predictions and, starting prior to image onset, transmits them to posterior cortex in the alpha band (*33*). Ketamine blocks predictions by reducing frontal alpha power to the same low level prior to all images (reducing SNR), leading to undifferentiated feedback. Overall, it suggests that NMDARs normally sharpen representations of predictions in frontal cortex, to enable PC. The data are less supportive of the classical view of sensation, with its reliance on feedforward processing to

reconstruct images, because one might have expected little systematic difference in behavioral and neural measures for different predictive conditions.

This has broader clinical and scientific relevance. Ketamine is a promising treatment for depression (*34*), and possible mechanisms include blocking lateral habenula bursting activity (*35*) and activating the mammalian target of rapamycin (mTOR) signaling pathway (*36*). Our results point to an additional antidepressant mechanism for ketamine. Depression is associated with negative predictions about upcoming personal events (*37*, *38*). By blocking predictions, ketamine may reduce the negative (*39*, *40*) bias, to ameliorate depressive symptoms (*41*). Finally, it has been proposed that generative models create virtual realities that support conscious experience (*3*, *4*, *10*). That subjects' predictions could be blocked, and error signaling reduced, without impairing consciousness imposes constraints on PC as a theory of consciousness.

References and Notes:

1. M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex. *Nat Neurosci*. **2**, 1019–25 (1999).

2. T. Serre, A. Oliva, T. Poggio, A feedforward architecture accounts for rapid categorization. *PNAS*. **104**, 6424–6429 (2007).

3. J. A. Hobson, K. J. Friston, Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*. **98**, 82–98 (2012).

4. A. K. Seth, K. Suzuki, H. D. Critchley, An Interoceptive Predictive Coding Model of Conscious Presence. *Front. Psychol.* **2** (2012), doi:10.3389/fpsyg.2011.00395.

5. P. Dayan, G. E. Hinton, R. M. Neal, R. S. Zemel, The Helmholtz machine. *Neural Comput*. **7**, 889–904 (1995).

6. M. W. Spratling, A review of predictive coding algorithms. *Brain Cogn*. **112**, 92–97 (2017).

7. R. P. Rao, D. H. Ballard, Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci*. **2**, 79–87 (1999).

8. D. Mumford, On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol Cybern*. **66**, 241–51 (1992).

9. K. Friston, The free-energy principle: a unified brain theory? *Nat Rev Neurosci*. **11**, 127–38 (2010).

10. J. A. Hobson, K. J. Friston, Consciousness, dreams, and inference: The Cartesian theatre revisited. *Journal of Consciousness Studies*. **21**, 6–32 (2014).

11. E. B. Issa, C. F. Cadieu, J. J. DiCarlo, Neural dynamics at successive stages of the ventral visual stream are consistent with hierarchical error signals. *Elife*. **7** (2018), doi:10.7554/eLife.42870.

12. N. Gordon, R. Koenig-Robert, N. Tsuchiya, J. J. van Boxtel, J. Hohwy, Neural markers of predictive coding under perceptual uncertainty revealed with Hierarchical Frequency Tagging. *Elife*. **6** (2017), doi:10.7554/eLife.22749.

13. T. Egner, J. M. Monti, C. Summerfield, Expectation and Surprise Determine Neural Population Responses in the Ventral Visual Stream. *J. Neurosci.* **30**, 16601–16608 (2010).

14. H. Homayoun, B. Moghaddam, NMDA receptor hypofunction produces opposite effects on prefrontal cortex interneurons and pyramidal neurons. *J Neurosci*. **27**, 11496–500 (2007).

15. H. Homayoun, B. Moghaddam, Orbitofrontal cortex neurons as a common target for classic and glutamatergic antipsychotic drugs. *Proc Natl Acad Sci U S A*. **105**, 18041–6 (2008).

16. R. E. Rosch, R. Auksztulewicz, P. D. Leung, K. J. Friston, T. Baldeweg, Selective Prefrontal Disinhibition in a Roving Auditory Oddball Paradigm Under N-Methyl-D-Aspartate Receptor Blockade. *Biol Psychiatry Cogn Neurosci Neuroimaging*. **4**, 140–150 (2019).

17. J. D. Murray, A. Anticevic, M. Gancsos, M. Ichinose, P. R. Corlett, J. H. Krystal, X. J. Wang, Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model. *Cereb Cortex*. **24**, 859–72 (2014).

18. A. M. Rosier, L. Arckens, G. A. Orban, F. Vandesande, Laminar distribution of NMDA receptors in cat and monkey visual cortex visualized by [3H]-MK-801 binding. *J Comp Neurol*. **335**, 369–80 (1993).

19. C. F. Zorumski, Y. Izumi, S. Mennerick, Ketamine: NMDA Receptors and Beyond. *J. Neurosci.* **36**, 11158–11164 (2016).

20. D. C. Javitt, M. Steinschneider, C. E. Schroeder, J. C. Arezzo, Role of cortical N-methyl-D-aspartate receptors in auditory sensory memory and mismatch negativity generation: implications for schizophrenia. *Proc Natl Acad Sci U S A*. **93**, 11962–7 (1996).

21. A. Schmidt, R. Bachmann, M. Kometer, P. A. Csomor, K. E. Stephan, E. Seifritz, F. X. Vollenweider, Mismatch negativity encoding of prediction errors predicts S-ketamine-induced cognitive impairments. *Neuropsychopharmacology*. **37**, 865–875 (2012).

22. P. R. Corlett, G. D. Honey, P. C. Fletcher, Prediction error, ketamine and psychosis: An updated model. *J Psychopharmacol*. **30**, 1145–1155 (2016).

23. X. Chen, S. Shu, D. A. Bayliss, HCN1 channel subunits are a molecular substrate for hypnotic actions of ketamine. *J. Neurosci.* **29**, 600–609 (2009).

24. G. Lupyan, S. L. Thompson-Schill, The evocative power of words: activation of concepts by verbal and nonverbal means. *J Exp Psychol Gen*. **141**, 170–86 (2012).

25. T. V. Wiecki, I. Sofer, M. J. Frank, HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Front Neuroinform*. **7**, 14 (2013).

26. T. L. Griffiths, J. B. Tenenbaum, Structure and strength in causal induction. *Cogn Psychol*. **51**, 334–84 (2005).

27. O. Jensen, A. Mazaheri, Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Front Hum Neurosci*. **4**, 186 (2010).

28. J. Lange, R. Oostenveld, P. Fries, Reduced occipital alpha power indexes enhanced excitability rather than improved visual perception. *J Neurosci*. **33**, 3212–20 (2013).

29. L. Ma, K. Skoblenick, J. K. Seamans, S. Everling, Ketamine-Induced Changes in the Signal and Noise of Rule Representation in Working Memory by Lateral Prefrontal Neurons. *J. Neurosci.* **35**, 11612–11622 (2015).

30. K. Skoblenick, S. Everling, NMDA antagonist ketamine reduces task selectivity in macaque dorsolateral prefrontal neurons and impairs performance of randomly interleaved prosaccades and antisaccades. *J. Neurosci.* **32**, 12018–12027 (2012).

31. M. W. Self, R. N. Kooijmans, H. Super, V. A. Lamme, P. R. Roelfsema, Different glutamate receptors convey feedforward and recurrent processing in macaque V1. *Proc Natl Acad Sci U S A*. **109**, 11031–6 (2012).

32. P. E. Vlisides, T. Bel-Bahar, U. Lee, D. Li, H. Kim, E. Janke, V. Tarnal, A. B. Pichurko, A. M. McKinney, B. S. Kunkler, P. Picton, G. A. Mashour, Neurophysiologic Correlates of Ketamine Sedation and Anesthesia: A High-density Electroencephalography Study in Healthy Volunteers. *Anesthes*. **127**, 58–69 (2017).

33. A. Alamia, R. VanRullen, Alpha oscillations and traveling waves: Signatures of predictive coding? *PLoS Biol*. **17**, e3000487 (2019).

34. M. Aan Het Rot, C. A. Zarate Jr., D. S. Charney, S. J. Mathew, Ketamine for depression: where do we go from here? *Biol Psychiatry*. **72**, 537–47 (2012).

35. Y. Yang, Y. Cui, K. Sang, Y. Dong, Z. Ni, S. Ma, H. Hu, Ketamine blocks bursting in the lateral habenula to rapidly relieve depression. *Nature*. **554**, 317–322 (2018).

36. N. Li, B. Lee, R.-J. Liu, M. Banasr, J. M. Dwyer, M. Iwata, X.-Y. Li, G. Aghajanian, R. S. Duman, mTOR-dependent synapse formation underlies the rapid antidepressant effects of NMDA antagonists. *Science*. **329**, 959–964 (2010).

37. D. R. Strunk, H. Lopez, R. J. DeRubeis, Depressive symptoms are associated with unrealistic negative predictions of future life events. *Behav Res Ther*. **44**, 861–882 (2006).

38. L. B. Alloy, A. H. Ahrens, Depression and pessimism for the future: biased use of statistically relevant information in predictions for self versus others. *J Pers Soc Psychol*. **52**, 366–378 (1987).

39. G. E. Bruder, J. P. Sedoruk, J. W. Stewart, P. J. McGrath, F. M. Quitkin, C. E. Tenke, EEG alpha measures predict therapeutic response to an SSRI antidepressant: Pre and post treatment findings. *Biol Psychiatry*. **63**, 1171–1177 (2008).

40. S. D. Muthukumaraswamy, A. D. Shaw, L. E. Jackson, J. Hall, R. Moran, N. Saxena, Evidence that Subanesthetic Doses of Ketamine Cause Sustained Disruptions of NMDA and AMPA-Mediated Frontoparietal Connectivity in Humans. *J Neurosci*. **35**, 11694–706 (2015).

41. T. Lyons, R. L. Carhart-Harris, More Realistic Forecasting of Future Life Events After Psilocybin for Treatment-Resistant Depression. *Front. Psychol.* **9** (2018), doi:10.3389/fpsyg.2018.01721.

Acknowledgments: We thank G. Lupyan, R.A. Pearce, B.R. Postle, and J. Samaha for useful discussions.

Funding: Y.S. supported by NIH grant R01MH110311. R.S. supported by NIH grant K23AG055700 and R01AG063849-01.

Author contributions: S.M., M.R., N.K., J.P., R.S. and Y.S. designed study; S.M., C.C., S.T., D.P., W.F., R.S., and Y.S. performed research; S.M., M.A., C.C., S.T., J.A., R.S. and Y.S. analyzed data; S.M., R.S. and Y.S. wrote paper; S.M., M.A., C.C., S.T., M.R., N.K., J.P, D.P., W.F., J.A., R.S., and Y.S. edited paper.

Competing interests: Authors declare no competing interests.

Data and materials availability: All data and code available upon reasonable request.

Supplementary Materials:

Materials and Methods

Figures S1-S5

Tables S1

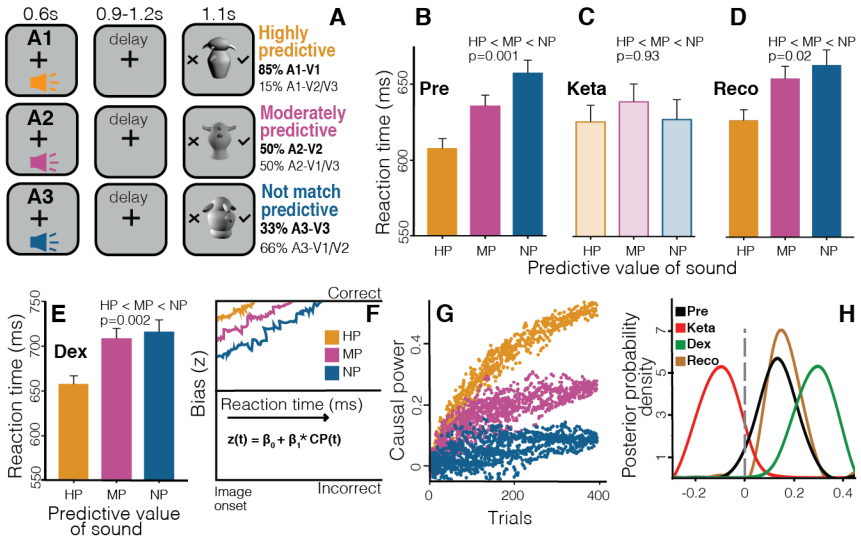

Fig. 1. Ketamine blocked fast RTs to predictive sounds. (A) We manipulated subjects' predictions by changing the probability of an image appearing after its associated sound in an audio-visual delayed match-to-sample task. Population RT (+SE) of 11 subjects (B) before (Pre), (C) under (Keta), and (D) after recovery (Reco) from ketamine. (E) Population RT (+SE) of 13 subjects under dexmedetomidine (DEX). (F) HDDM of RTs. Bias (z) calculated for each trial (t) using causal power (CP). β_1 determines relationship between z and CP. (G) Population CP values across time for pre-ketamine (Pre) testing. (H) Posterior probability density of β_1 for different drug conditions.

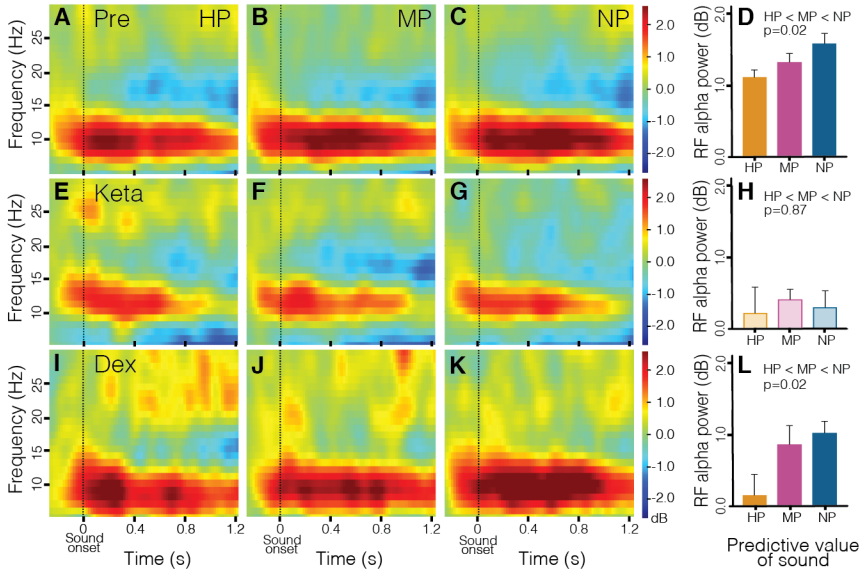

Fig. 2. Ketamine blocked correlation between prediction strength and right frontal alpha power. Population time-frequency decomposition of right frontal electrode cluster (RF) before drug administration (Pre; A-C), under ketamine (Keta; E-G), and under dexmedetomidine (Dex; I-K), for highly predictive (HP; A, E and I), moderately predictive (MP; B, F and J), and not-match predictive (NP; C, G and K) sounds. Power calculated in 0.55 s sliding windows, with window at 0 s representing interval -0.275 s to +0.275 s. Population average RF alpha power in delay period (after 0.6 s) for (D) Pre, (H) Keta and (L) Dex.

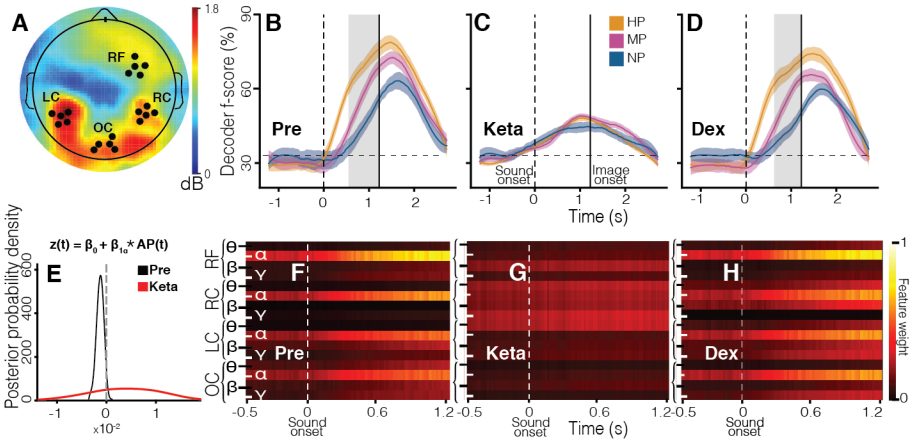

Fig. 3. Ketamine blocked decoding of predictions. (A) Electrode clusters in right frontal (RF), right central (RC), left central (LC), and occipital (OC) cortex showing significant modulation in delay period alpha power. Average across all trials (HP, MP and NP). F score using support vector machine (SVM) to decode predictive value (highly predictive, HP; moderately predictive, MP; or not-match predictive, NP) of sound, based on time-frequency power spectrum (B) before drug administration (Pre), (B) under ketamine (Keta), and (C) under dexmedetomidine (Dex). Dashed vertical line denotes sound onset. Solid vertical line denotes earliest possible image onset. Dashed horizontal line signifies level of decoding expected by chance. Gray shaded areas indicate significant zone of separation between HP, MP and NP, prior to image onset. (E) Posterior probability density of $\beta_{1\alpha}$. Bias (z) calculated for each trial (t) using alpha power (AP). $\beta_{1\alpha}$ determines relationship between z and AP. Feature weights from SVM decoder for (F) Pre, (G) Keta, and (H) Dex. θ, α, β, and γ indicate theta (5-7 Hz), alpha (8-14 Hz), beta (15-30 Hz), and gamma (30-45 Hz) bands respectively.

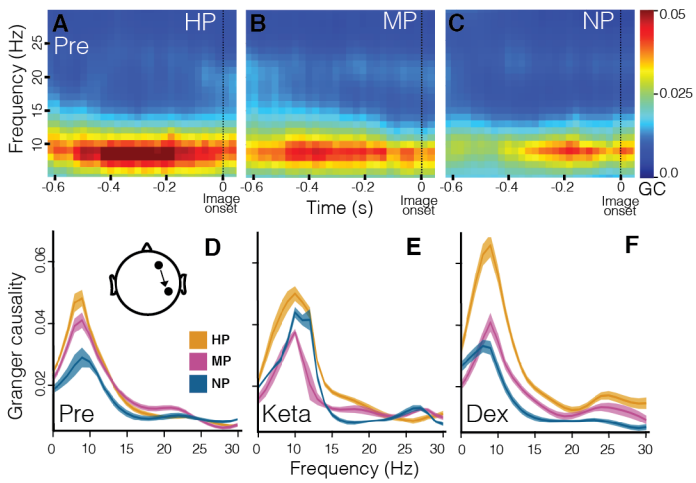

Fig. 4. Ketamine perturbed correlation between prediction strength and alpha feedback. Time-frequency plots of Granger causality from right frontal to right central cluster before drug administration (Pre; A-C), for HP (A), MP (B), and NP (C) sounds. Granger causality calculated in 0.55 s sliding windows, with window at 0 s representing interval -0.275 s to +0.275 s. Plots aligned to image onset. Population average Granger causal influence during delay period for (D) Pre, (E) Keta, and (F) Dex.

Supplementary Materials for

Receptors, circuits and neural dynamics in prediction

Sounak Mohanta, Mohsen Afrasiabi, Cameron Casey, Sean Tanabe, Michelle J. Redinbaugh, Niranjan A. Kambi, Jessica M. Phillips, Daniel Polyakov, William Filbey, Joseph L. Austerweil, Robert D. Sanders, Yuri B. Saalmann

Correspondence to: saalmann@wisc.edu or robert.sanders@wisc.edu

This PDF file includes:

Materials and Methods

Participants

The University of Wisconsin-Madison Health Sciences and Social Sciences Institutional Review Boards (IRBs) approved experiments. 29 participants (14 female) performed the psychophysics predictive coding experiment. We excluded data from four subjects as their performance accuracy was below 50%. 17 additional participants (six female) for dexmedetomidine and 11 participants (five female) for ketamine took part in the pharmacology predictive coding experiments. Four participants (from 17 dexmedetomidine participants) were excluded for low accuracy (accuracy less than 50%). Participants who performed the psychophysics predictive coding experiment did not take part in the pharmacology experiments. All 11 participants who performed ketamine experiments had to participate in the dexmedetomidine experiments first as per IRB requirements.

Stimuli

For our psychophysics experiments, we used biomorphic visual stimuli from Michael Tarr's lab (http://wiki.cnbc.cmu.edu/Novel_Objects). These are known as greebles. Fig. 1A shows examples presented to participants. We used three gray-scale greebles for each psychophysics session, and each greeble was personified with a name. We used novel sounds (trisyllabic nonsense words) for the greeble names – e.g., "Tilado", "Paluti", and "Kagotu" – from Saffran et al. (*1*). The sounds were generated using the Damayanti voice in the "text to speech" platform of an Apple MacBook. To avoid differences in the

salience of stimuli, greeble images have similar size (13 degrees of visual angle in height and 8 degrees in width), number of extensions and mean contrast, and greeble names have the same number of syllables and sound level (80 dB SPL).

For the pharmacology experiments, we generated three new triplets of greebles. To control for saliency, each participant rated greeble salience for each of four (three plus one from the psychophysics experiment) triplets, i.e., the participant identified whether any of the greebles in a triplet stand out compared to the other two greebles from the same triplet. We proceeded to use triplets which the participant rated all three greebles as being equally salient. We then named each of these greebles with a new trisyllabic nonsense word.

Audio-visual delayed match-to-sample task

Each trial of the task involves the sequential presentation of a sound (trisyllabic nonsense word) followed by a greeble image. We refer to stimuli using the following notation: A1, A2 and A3 correspond to each of the three sounds used (A for auditory) and V1, V2 and V3 correspond to each of the greebles used (V for visual). Using this notation, audio-visual stimulus sequences containing the matching name and greeble are A1-V1, A2-V2 and A3-V3. Audio-visual stimulus sequences containing a non-matching name and greeble are A1-V2, A1-V3, A2-V1, A2-V3, A3-V1 and A3-V2. We pseudo-randomized names for greebles (i.e., matching sounds and images) across subjects.

Learning Phase. During the first phase of the task, participants learn the association between the sounds and images (i.e., names of the greebles) through trial-and-error, by

performing a match/non-match (M/NM) task. This phase is called the "learning" phase. Each trial starts with blank blue screen (R=35, G=117, B=208, 200ms duration; shown as gray in Fig. 1A). After that, a black fixation cross (size 1.16 degrees of visual angle; jittered 200-400 ms) is presented followed by a sound, a greeble name voiced by the computer (600 ms duration). After a jittered delay period (900-1,200 ms duration), a greeble image (until a participant responds or 1,100 ms duration, whichever is earliest) was presented on the monitor screen, as well as two symbols (√ and **X**) to the left and right of the greeble (9.3 degrees of visual angle from screen center). These symbols indicated participants' two response options: match (√) or non-match (**X**). The symbol location, left or right of the greeble image, corresponded to the left or right response button, respectively: left and right arrow keys of a computer keyboard in the psychophysics experiments; and left and right buttons of a mouse in the pharmacology experiments. We randomly varied the symbols' locations relative to the greeble image to minimize motor preparation (i.e., on some trials, a match response required a left button press and, on other trials, a match response required a right button press). In the learning phase, each greeble name and image had 33% probability of appearing in any given trial. This is to prevent subjects from developing any differential predictions about the greebles due to greeble name or image frequency, during the learning phase.

To address possible same-different biases, e.g., quicker reaction times (RTs) for match trials (*2*) we introduced a control called "inversion trials" in psychophysics experiments, to minimize the expectation of M/NM trials which, in itself, might otherwise contribute to participants' responses. In these inversion trials, participants had to respond whether the

greeble image presented on screen is inverted (the appropriate response button, left/right, indicated on screen by the left/right location of a red arrow pointing down) or upright (yellow arrow pointing up). Participants did not know the type of trial in advance; the trial type was only revealed by the symbols to the left and right of screen center at the onset of the greeble image (i.e., √ and **X** signal M/NM trials, whereas downward red arrow and upward yellow arrow signal inversion trials). 50% of the total number of trials in the learning phase were inversion trials and the rest were M/NM trials. Because participants cannot specifically prepare in advance for M/NM trials due to the random presentation of inversion and M/NM trials, there should be minimal confounding of RTs with a bias towards match responses.

Testing phase. Once participants show above 80% accuracy for the M/NM trials in the learning phase of the task, they move on to the "testing phase" (1,000 trials for the psychophysics experiments; Fig. 1A). During the testing phase, we manipulated predictions by changing the probability of a greeble appearing after its learnt name. This probability is different for each greeble name and image. That is, in the testing phase, when a participant hears A1, there is 85% chance of V1 being shown (highly predictive; HP); when a participant hears A2, there is 50% chance of V2 being shown (moderately predictive; MP); and when a participant hears A3, there is a 33% chance of V3 being shown (not-match predictive; NP). This allows participants to make stronger predictions about the identity of the upcoming visual image after hearing A1, than after hearing A2 or A3, for instance.

Inversion trials also consisted of half the total trials in the testing phase of the psychophysics experiments. We randomly presented all the trial types (M/NM trials and inversion trials) to the participants. The testing phase of the task had approximately equal match and non-match trials to avoid response bias. The testing phase also had an approximately equal number of trials for each greeble image, and its corresponding name had approximately equal probability of being voiced, to control for stimulus familiarity.

Causal Strength and Transitional Probability

We quantified the relationship between a sound (name) and its paired image (greeble) using the strength of causal induction: given a candidate cause *C* (sound) how likely is the effect *E* (i.e., how likely is it followed by its paired image). We will represent variables *C* and *E* with upper case letters, and their instantiations with lower case letters. Hence, $C = c+ / E = e+$ indicates that the cause/effect is present, and $C = c- / E = e-$ indicates that the cause/effect is absent (for brevity, we will shorten variables equal to outcomes, such as $C = c+$ or $C = c-$ as simply $c+$ or $c-$, respectively). The evidence for a relationship can be encoded as a 2 X 2 contingency table for each sound, as in Table S1 (black letters), where $\mathrm{N(c+, e+)}$ represents the number of trials in which the effect occurs in the presence of the cause, $\mathrm{N(c-, e+)}$ represents the number of trials in which the effect occurs in the absence of the cause and so on. Applied to our study, e.g., *C* could be hearing sound A1, and *E* viewing the paired greeble V1. For this case, $\mathrm{N(c+, e+)}$ would be the number of trials V1 follows A1; whereas $\mathrm{N(c-, e+)}$ would be the number of trials V1 follows A2 or A3. The full contingency table for the "Highly Predictive" auditory cue A1 and its paired greeble V1 is shown in Table S1 (in green letters). There

are analogous contingency tables for the other two auditory cues and their paired greebles.

Based on these contingency tables, we calculated three different measures of causal relationship (ΔP, causal power and causal support) for each trial. ΔP and causal power assume that *C* causes *E*. ΔP reflects how the probability of *E* changes as a consequence of the occurrence of the *C*. Causal power corresponds to the probability that an effect *E* happened because of cause *C* in the absence of all other causes. Whereas causal support evaluates whether or not a causal relationship actually exists and calculates the strength of that relationship. To do this, causal support estimates the evidence for a graphical model with a link between *C* and *E* against one without a link. For example, let us consider the graphs denoted by Graph 0 and Graph 1 in Fig. S1D (adapted from (*3*)). There are three variables in each graph: cause *C*, effect *E* and background cause B. In Graph 0, *B* causes *E*, but *C* has no relationship to either *B* or *E*. In Graph 1, both *B* and *C* cause *E*. While calculating ΔP and causal power Graph 1 is assumed, whereas causal support compares the structure of Graph 1 to that of Graph 0. Causal support is defined as the evidence provided from data D in favor of Graph 1, P(D | Graph 1), over Graph 0, P(D | Graph 0), which can be calculated by the following equation:

$$Causal\ support = \log \frac{\mathrm{P(D|Graph\ 1)}}{\mathrm{P(D|Graph\ 0)}} \qquad (1)$$

We calculated causal support using freely available Matlab code from (*3*). ΔP and causal power were calculated using the following formulas:

$$\Delta P = \frac{N(c+,e+)}{N(c+,e+) + N(c+,e-)} - \frac{N(c-,e+)}{N(c-,e+) + N(c-,e-)} = P(e+|c+) - P(e+|c-) \qquad (2)$$

$$Causal\ power = \frac{\Delta P}{1 - P(e+|c-)} \qquad (3)$$

We compared these three measures of causal relationship with the transitional probability (i.e., a comparison between causation and correlation). We calculated the transitional probability of each greeble (V) given prior presentation of its paired sound (A), using the equation below:

$$\text{Transitional probability} = \frac{N(c+,e+)}{N(c+)} \qquad (4)$$

For each subject, we used the causal relationship value of each condition (HP, MP and NP) at the end of a testing phase as the starting values of the next testing phase. For example, the starting values of causal relationship for the "under drug" testing phase were equal to the causal relationship values at the end of the "pre/baseline" testing phase. Similarly, the starting values for the "after recovery" testing phase were equal to the causal relationship values at the end of the "under drug" phase (Fig. S2A and B). We tested if subjects (i) regained access to already learned and stored predictive information, after they recovered from ketamine dosing, or (ii) re-learned the predictive information. We mimicked hypothesis (ii) by forcing the starting values of causal relationship (causal power here) "after recovery" to be zero (Fig. S2C) instead of starting values equal to the causal relationship at the end of the "under drug" phase (hypothesis (i); Fig. S2B).

Pharmacology Experiments

To manipulate participants' predictions, we administered two drugs, ketamine and dexmedetomidine, each on a separate day: dexmedetomidine on the first day, and ketamine on the second day, with at least one month intervening. This fixed order was IRB-imposed in their consideration of safety profiles of the different medications

(registered on NCT03284307). All subjects were healthy and aged between 18-40 years old without contraindication to study drugs. We also acquired EEG data throughout pharmacology experiments (see section "EEG Recording" below), to measure electrophysiological activity during predictive coding. A typical pharmacology experiment consisted of three segments: (a) pre-drug baseline, (b) under drug influence (dexmedetomidine or ketamine), and (c) after recovery. During the pre-drug baseline, participants performed the learning phase (200 trials), then the first testing phase (400 trials). Under drug influence, participants performed the second testing phase (200 trials). Under ketamine, participants also performed a third testing phase (200 trials; see ketamine dosing for details). After recovery, they performed the last testing phase (400 trials). Due to a protocol-limited maximum time under drug influence, and the need to acquire sufficient M/NM trials for EEG analyses, the pharmacology experiments did not include inversion trials. All other aspects of the task in pharmacology experiments were the same as that in psychophysics experiments. For each of the two pharmacology experiments involving a particular participant, we used three new greeble name and image pairs to rule out any possible contribution of long-term memory.

Dexmedetomidine Dosing

We intravenously administered a 0.5 mcg/kg bolus over 10 minutes, followed by 0.5 mcg/kg/h infusion. Participants performed the testing phase (under drug influence) during this infusion time, corresponding to stable drug levels according to the pharmacokinetic model for dexmedetomidine by Hannivoort et al. (*4*). We targeted a plasma concentration of dexmedetomidine that is associated with mild sedation (modified observer's

assessment of alertness/sedation (OAA/S) of 4 (*5*)) to control for non-specific sedative effects, including hyperpolarization-activated cyclic nucleotide channel (HCN-1) effects (*6*). The actual sedation achieved was on average slightly deeper than anticipated (modified OAA/S median 3, IQR 2) with a mean plasma concentration of 0.8 (SD 0.33) ng/ml (*5*).

Ketamine Dosing

In initial experiments, we tested two doses of ketamine to target the lowest plasma concentration that would modulate NMDARs in the relevant concentration range (<1microMolar (*7*)). The first dose corresponded to intravenously administered 0.25 mg/kg ketamine over 10 minutes, followed by 30 mg/h infusion, corresponding to 0.4 microMolar. Prior to testing subjects, lack of nystagmus and visual disturbance was confirmed in all participants. Participants performed the testing phase (under drug influence) during this infusion time, corresponding to stable drug levels according to the pharmacokinetic model for ketamine by Domino et al. (*8*, *9*). A second dose was tested with a second bolus of 0.25 mg/kg ketamine over 10 minutes, followed again by 30 mg/h infusion. Testing again was completed once stable plasma concentrations of approximately 0.8 microMolar were achieved. Ketamine blocked predictions at this second level of ketamine dosing, equating to a minimum plasma concentration of 0.2 μg/ml; range tested: 0.2-0.3 μg/ml. As we found the effective ketamine dose to be 0.2 μg/ml in our first seven subjects, we targeted that plasma concentration for our last four subjects. Two subjects were excluded from “after recovery” testing due to vomiting.

Monitoring

Subjects were monitored during drug exposure according to the American Society of Anesthesiologists guidelines, including electrocardiogram, blood pressure and oxygen saturation. We monitored arousal level using the modified OAA/S scale (*10*).

EEG Recording

We performed high-density EEG recordings using a 256 channel system (including NA 300 amplifier; Electrical Geodesics, Inc., Eugene, OR). After applying the EEG cap with conductive gel (ECI Electro-Gel), we adjusted electrodes so that the impedance of each electrode was within 0-50 kiloohms. We checked electrode impedance before the experiment started, and again before drug administration. Using Net Station, we sampled EEG signals at 250 Hz and, off-line, bandpass filtered between 0.1 Hz and 45 Hz.

EEG Preprocessing

We combined pre-drug baseline data from both ketamine and dexmedetomidine experiments (baseline RT data showed similar results for both drugs) but "under drug" analyses were performed separately for each drug. We performed offline preprocessing and analysis using EEGLAB (*11*). First, we extracted data epochs -1,500 ms to 3,000 ms relative to the onset of the sound and -3000 ms to 800 ms relative to the onset of the image, for each trial. We then visually inspected each epoch and excluded noisy trials (around 5% of the total trials). Next, we performed Independent Component Analysis (ICA) using built-in functions of EEGLAB (pop_runica.m) and removed noisy components through visual inspection. We excluded two subjects from further analysis – one ketamine

subject and one dexmedetomidine subject – due to very noisy EEG data, which after cleaning left too few trials for analysis (conditions with <10 trials). Finally, we performed channel interpolation (EEGLAB function, eeg_interp.m, spherical interpolation) and re-referenced to the average reference.

Time-Frequency Decomposition

To investigate changes in EEG spectral content, we performed time-frequency decompositions of the preprocessed data using Morlet wavelets, whose frequency ranged from 5 Hz to 45 Hz in 40 linearly spaced steps. Power for each time-frequency point is the absolute value of the resulting complex signal. We dB normalized power (dB power = 10*log10[power/baseline]) to the baseline of -400 ms to 0 ms relative to the onset of the sound. For drift-diffusion model analysis, we calculated spectral power and performed divisive baseline correction for each trial.

Electrode Selection

We used a data-driven approach, orthogonal to the effect of interest, to select the electrodes of interest based on the task EEG data. In the first step, we averaged power across all electrodes (aligned to sound onset) and all sounds/greeble names. This revealed increased alpha power (8-14 Hz) during the delay period compared to baseline (Fig. S3A). In the second step, we selected the electrode clusters that showed significant change in alpha power during the 600 ms to 1,225 ms time window post-sound onset, compared to baseline. This time window ensured that our delay period activity was devoid of any stimulus-driven response, either from the sound or image. After cluster-based

multiple comparisons correction (*6*), four different clusters showed significant modulation in alpha power during the delay period (Fig. 3A): (i) right frontal (RF electrodes 4, 214, 215, 223, 224); (ii) left central (LC electrodes 65, 66, 71, 76, 77); (iii) right central (RC electrodes 163, 164, 173, 181, 182); and (iv) occipital (OC electrodes 117, 118, 119, 127, 129).

Alpha Power Calculation

For each condition, HP, MP and NP, we averaged alpha power over 600 to 1,225 ms (aligned to sound onset) and all five electrodes in a cluster, to calculate the average alpha power of each cluster. To best capture the delay period activity just prior to image onset, we also calculated mean alpha power between -625 to -275ms (as the wavelet window is centered around each time point, the power estimate before -625 ms and after -275 ms may contain auditory and visual stimulus-related responses, respectively) for each trial aligned to image onset. To link EEG spectral power to behavior using our drift-diffusion model analysis (see section HDDM), we calculated single-trial baseline-corrected (divisive normalization) alpha band power, aligned to image onset.

Granger Causality

We calculated spectral Granger causality (GC) between electrodes in the right frontal and right central clusters using the SIFT plugin (https://sccn.ucsd.edu/wiki/SIFT) of EEGLAB (*12*, *13*). GC was calculated on the image-aligned data to best capture delay period activity just before the image onset. For each cluster, in each subject, we used the electrode that showed the largest change in alpha power relative to baseline (based on

the electrode selection procedure above). First, we downsampled the data to 125 Hz. Second, we linearly detrended the data, then normalized it (divided the detrended data by its standard deviation). Next, we performed multivariate autoregressive modelling (MVAR) using a model order of 10, which corresponded to the first minimum Akaike information criterion value. Two models were fitted for each subject: one for the pre-drug baseline, and one under drug influence. For each sound (A1, A2 and A3), we calculated GC using a window length of 550 ms and a step size of 30 ms. We validated our models through checks for whiteness of residuals, percent consistency and model stability. All models were stable and had an average of 86% consistency. We calculated mean GC at each frequency (linear intervals) by averaging over the delay period (-625 ms to -275 ms; see section "Alpha Power Calculation" for justification of window length).

Hierarchical Drift-Diffusion Model (HDDM)

We used a drift-diffusion model (DDM) (*14*), where there are two possible choices (correct/incorrect responses of the match trials) in our predictive coding task. According to this model, decision-making involves the accumulation of evidence (drift process) from a starting point to one of two (upper or lower) thresholds, representing the choices. The accumulation rate is known as the drift rate, v; and the starting point can be biased towards one of the choices (in our study, by the predictive value of the sound), reflected in a bias parameter, z. We used HDDM software (http://ski.clps.brown.edu/hddm_docs/) (*15*) for hierarchical Bayesian estimation of the parameters of the drift-diffusion model. The hierarchical approach allows estimates of the group level and individual subject level parameters simultaneously. Particularly with fewer trials per condition, this method has

been shown to provide more reliable estimates of parameters and is less susceptible to outliers (*14*) than more traditional approaches to DDMs (*16*, *17*) .

To directly link the causal relationship between the sound and its paired greeble image to behavior and drift-diffusion parameters, we included the estimates of causal relationship and transitional probability as predictor variables of the bias, z, of the model. That is, we estimated posterior probability densities not only for basic model parameters, but also the degree to which these parameters are altered by variations in the psychophysical measures (ΔP, causal power, causal support and transitional probability). In these regressions, the bias parameter is given by, $z(t) = \beta_0 + \beta_1 CP(t)$, where CP is either ΔP, causal power, causal support or transitional probability, β_0 is the intercept, and t is the trial number. Here, the slope, β_1, is weighted by the value of the psychophysical measure on that trial. The regression across trials allows us to infer how the bias changes depending on the psychophysical measure. For example, if these psychophysical measures are positively correlated to bias, then increased causal strength or transitional probability will yield faster RTs. We fit four different versions of the model: (i) ΔP Model, where bias was estimated from the ΔP and updated after each trial; (ii) Causal Power Model, where bias was estimated from the causal power and updated after each trial; (iii) Causal Support Model, where bias was estimated from the causal support and updated after each trial; and (iv) Transitional Probability Model, where bias was estimated from the transitional probability and updated after each trial. We also modeled our data where drift rate varied according to causal strength or transitional probability. We used the Deviance Information Criterion (DIC) for model comparison (*18*). The DIC is a measure

of model fit (i.e., lack thereof) with a penalty for complexity (i.e., the number of parameters used to fit the model to the data) (*19*). Models with lower DIC are better models. Models where bias was manipulated instead of drift rate had significantly lower DIC. For the rest of our analyses, we used models where bias varied with causal strength or transitional probability and drift rate was kept constant at the default values, as this yielded the lowest DIC.

We used Markov Chain Monte Carlo chains with 20,000 samples and 5000 burn-in samples for estimating the posterior distributions of the model parameters. We assessed chain convergence by visually inspecting the autocorrelation distribution, as well as by using the Gelman-Rubin statistic, which compares between-chain and within-chain variance. This statistic was near 1.0 for the parameters, indicating that our sampling was sufficient for proper convergence. We analyzed parameters of the best model (model with lowest DIC) using Bayesian hypothesis testing, where the percentage of samples drawn from the posterior fall within a certain region (e.g., > 0). Posterior probabilities $\geq 95\%$ were considered significant. Please note that this value is not equivalent to p-values estimated by frequentist methods, but they can be coarsely interpreted in a similar manner.

All the model comparisons were estimated on the psychophysics data as these had the greatest number of trials per condition. This ensures robust estimation of the best model. The best-fitting model was then used to analyze data from different conditions: pre-drug baseline, under dexmedetomidine, under ketamine and recovery from ketamine. To directly link EEG spectral power to behavior and drift-diffusion parameters, we used the HDDM, but now included the right frontal cluster power estimate (aligned to image onset)

in the alpha band as the predictor variable of the bias, *z*, of the model; i.e., in the regression equation above, CP was now alpha power.

Event-related Potentials

We used ERPLAB (https://erpinfo.org/erplab) (*20*) to run event-related potential (ERP) analysis. First, we cleaned epoched data (aligned to the sound onset for auditory ERPs and to visual image onset for visual ERPs) using the pop_artmwppth.m function of ERPLAB with a moving window of 200 ms (2-3% of trials for each subject were excluded). We then averaged over trials to generate an average ERP for each subject. Based on previous literature (*21–24*), we chose channel 9 (Cz electrode) for auditory ERPs and channel 178 (T6) for visual ERPs. For auditory ERPs, we calculated the N200 response (negative peak amplitude between 200-300 ms) for each sound. We calculated the N170 response (negative peak amplitude between 150-200 ms) for the predicted (A1-V1) and not-predicted (A2-V1) image, V1. The predicted response was subtracted from the not-predicted response to calculate the mismatch negativity (MMN).

EEG Signal Decoding

We used the spectral power of EEG signals across four frequency bands, theta (5-7 Hz), alpha (8-14Hz), beta (15-30 Hz) and gamma (31-40 Hz), for the decoding analysis. We calculated the sum of squared absolute power in each frequency band for each electrode cluster (RF, LC, RC and OC), thus generating 16 features for each trial as the input dataset. For each 20 ms time bin, we trained a Support Vector Machine (SVM) model to classify the EEG data into three classes: highly predictive (HP), moderately predictive

(MP) and not-match predictive (NP). We denote $y_i(t) \in \{0,1,2\}$ as the identifier of the condition at time bin t, where 0, 1, and 2 denotes HP, MP and NP respectively. The SVM classifier is implemented by a nonlinear projection of the training data $\mathbf{x}(t)$ feature space $\mathcal{X}$ into a high dimensional feature space $\mathcal{F}$ using a kernel function ϕ. So with $\phi: \mathcal{X} \longrightarrow \mathcal{F}$ being the mapping kernel, the weight vector $\mathbf{w}$ can be expressed as a linear combination of the training trials and the kernel trick can be used to express the discriminant function as

$$y(\mathbf{x}(t); \zeta(t)) = \mathrm{a}^T(t)\phi_{\mathbf{x}}(t) + b(t) = \sum_{n=1}^{N} a_n(t)\varphi(\mathbf{x}_n(t), \mathbf{x}(t)) + b(t) \qquad (5)$$

where $\zeta(t) = \{\mathrm{a}(t), b(t)\}$ is the new parameter at time bin t with $\mathrm{a}(t)$ and $\mathrm{b}(t)$ as weights and biases of the mapped features space $\mathcal{F}$. We used the radial basis function (RBF) kernel that allows nonlinear decision boundary implementation in the input space. The RBF kernel holds the elements

$$\varphi\big(\mathbf{x}_n(t), \mathbf{x}(t)\big) = \exp(-\eta(t)\|\mathbf{x}_n(t) - \mathbf{x}(t)\|^2) \qquad (6)$$

where $\eta(t)$ is a tunable parameter at each time bin. Model hypermeters consisting of regularization penalty ($C(t)$) and $\eta(t)$ were selected by grid search through 10-fold cross validation. F-score at each time bin and for each label was calculated as

$$F - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (7)$$

where

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad (8)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \qquad (9)$$

As we penalized the mapped weights of the classifier at each time bin, we used normalized absolute values of the weights as a measure to deduce each feature's contribution to classify the outputs.

Statistical Analysis

We performed statistical analysis of trials from testing phases, and used the learning phase only to confirm that the participants learned the correct associations. To only include the trials where the causal power of HP, MP and NP trials have differentiated (Fig. 1G), we excluded the first 50 trials of the testing phase in psychophysics experiments and the pre-drug baseline. We used all available trials for the testing phase of "under drug" and "after recovery", as the causal power of HP, MP and NP trials were already differentiated from the beginning (Fig. S2, A and B). For RT analysis, we excluded RTs more/less than the mean +/- 3SD for each subject, and we used correct match trials. For delay period EEG analyses, we used both correct match and non-match trials. The data were analyzed using a linear mixed effect model with the predictive value of sound (HP, MP or NP; varying within subject) as the independent variable. We used contrast analysis to model the three categories of sounds. We used a linear (-1, 0, 1) contrast as our contrast of interest, and a quadratic (-1, 2, -1) contrast as the orthogonal contrast in the analysis. We used the same linear mixed effect model only changing the dependent variable with respect to hypotheses; i.e., RTs, accuracy (correct versus incorrect), average right frontal alpha power (aligned to sound onset), average GC at alpha frequency, or N200 amplitude in the auditory ERP. Our study conformed to the guidelines

set out by Ableson and Prentice (*25*) with regards to contrast analysis; i.e., we included the contrast of interest along with a paired, orthogonal contrast. Our statistical tests showed that contrasts of interest were significant while the orthogonal contrasts were not. To test if the MMN for the image is significantly less than zero, we used a paired t-test and FDR corrected for multiple comparisons. We used repeated measures ANOVA (Holm-Sidak corrected p-values) to test the significance of F-scores for the three differentially predictive conditions (HP, MP, NP) as well as the significance of each feature's contribution in output classification.

References

1. J. R. Saffran, R. N. Aslin, E. L. Newport, Statistical learning by 8-month-old infants. *Science*. **274**, 1926–8 (1996).

2. G. Lupyan, S. L. Thompson-Schill, The evocative power of words: activation of concepts by verbal and nonverbal means. *J Exp Psychol Gen*. **141**, 170–86 (2012).

3. T. L. Griffiths, J. B. Tenenbaum, Structure and strength in causal induction. *Cogn Psychol*. **51**, 334–84 (2005).

4. L. N. Hannivoort, D. J. Eleveld, J. H. Proost, K. M. E. M. Reyntjens, A. R. Absalom, H. E. M. Vereecke, M. M. R. F. Struys, Development of an Optimized Pharmacokinetic Model of Dexmedetomidine Using Target-controlled Infusion in Healthy Volunteers. *Anesthesiology*. **123**, 357–367 (2015).

5. P. J. Colin, L. N. Hannivoort, D. J. Eleveld, K. M. E. M. Reyntjens, A. R. Absalom, H. E. M. Vereecke, M. M. R. F. Struys, Dexmedetomidine pharmacokinetic-pharmacodynamic modelling in healthy volunteers: 1. Influence of arousal on bispectral index and sedation. *Br J Anaesth*. **119**, 200–210 (2017).

6. Y. Yang, Q. Meng, X. Pan, Z. Xia, X. Chen, Dexmedetomidine produced analgesic effect via inhibition of HCN currents. *Eur. J. Pharmacol.* **740**, 560–564 (2014).

7. P. J. Morris, R. Moaddel, P. Zanos, C. E. Moore, T. Gould, C. A. Zarate, C. J. Thomas, Synthesis and N-Methyl-D-aspartate (NMDA) Receptor Activity of Ketamine Metabolites. *Org Lett*. **19**, 4572–4575 (2017).

8. E. F. Domino, S. E. Domino, R. E. Smith, L. E. Domino, J. R. Goulet, K. E. Domino, E. K. Zsigmond, Ketamine kinetics in unmedicated and diazepam-premedicated subjects. *Clin. Pharmacol. Ther.* **36**, 645–653 (1984).

9. E. F. Domino, E. K. Zsigmond, L. E. Domino, K. E. Domino, S. P. Kothary, S. E. Domino, Plasma levels of ketamine and two of its metabolites in surgical patients using a gas chromatographic mass fragmentographic assay. *Anesth. Analg.* **61**, 87–92 (1982).

10. D. A. P. Chernik, D. P. Gillings, H. P. Laine, J. M. Hendler, J. M. Silver, A. B. P. Davidson, E. M. P. Schwam, J. L. P. Siegel, Validity and Reliability of the Observer's: Assessment of Alertness/Sedation Scale: Study with: Intravenous Midazolam. *Journal of Clinical Psychopharmacology*. **10**, 244–251 (1990).

11. A. Delorme, S. Makeig, EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods*. **134**, 9–21 (2004).

12. A. Delorme, T. Mullen, C. Kothe, Z. Akalin Acar, N. Bigdely-Shamlo, A. Vankov, S. Makeig, EEGLAB, SIFT, NFT, BCILAB, and ERICA: New Tools for Advanced EEG Processing. *Computational Intelligence and Neuroscience*. **2011**, 12 (2011).

13. T. Mullen, thesis, UC San Diego (2014).

14. R. Ratcliff, G. McKoon, The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput*. **20**, 873–922 (2008).

15. T. V. Wiecki, I. Sofer, M. J. Frank, HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Front Neuroinform*. **7**, 14 (2013).

16. A. Voss, J. Voss, A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology*. **52**, 1–9 (2008).

17. J. Vandekerckhove, F. Tuerlinckx, Diffusion model analysis with MATLAB: a DMAT primer. *Behavior research methods*. **40**, 61–72 (2008).

18. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin, *Bayesian Data Analysis* (Chapman and Hall/CRC, 2013; https://www.taylorfrancis.com/books/9780429113079).

19. D. J. Spiegelhalter, N. G. Best, B. P. Carlin, A. V. D. Linde, Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. **64**, 583–639 (2002).

20. J. Lopez-Calderon, S. J. Luck, ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Front. Hum. Neurosci.* **8** (2014), doi:10.3389/fnhum.2014.00213.

21. B. Rossion, I. Gauthier, M. J. Tarr, P. Despland, R. Bruyer, S. Linotte, M. Crommelinck, The N170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: an electrophysiological account of face-specific processes in the human brain. *Neuroreport*. **11**, 69–74 (2000).

22. B. Rossion, I. Gauthier, V. Goffaux, M. J. Tarr, M. Crommelinck, Expertise training with novel objects leads to left-lateralized facelike electrophysiological responses. *Psychol Sci*. **13**, 250–257 (2002).

23. I. Winkler, S. Denham, C. Escera, in *Encyclopedia of Computational Neuroscience*, D. Jaeger, R. Jung, Eds. (Springer, New York, NY, 2013; https://doi.org/10.1007/978-1-4614-7320-6_99-1), pp. 1–29.

24. R. Ceponiene, T. Lepistö, M. Soininen, E. Aronen, P. Alku, R. Näätänen, Event-related potentials associated with sound discrimination versus novelty detection in children. *Psychophysiology*. **41**, 130–141 (2004).

25. R. P. Abelson, D. A. Prentice, Contrast tests of interaction hypothesis. *Psychological Methods*. **2**, 315–328 (1997).

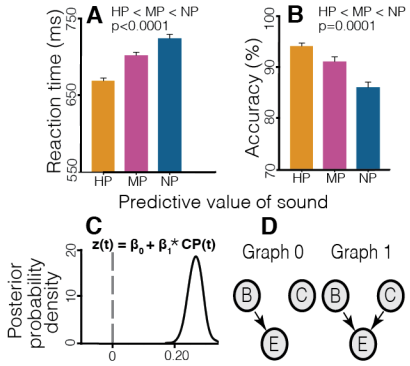

Fig. S1. Fast RTs to predictive sounds not due to speed-accuracy trade off. Population average (25 subjects from psychophysics experiment) (A) RT and (B) accuracy for highly predictive (HP), moderately predictive (MP), and not-match predictive (NP) sounds. (C) Posterior probability density of β_1 of hierarchical drift diffusion model from psychophysics experiment. (D) Directed graphs for calculating causal support. In graph 0, B causes E, but C has no relationship to either B or E. In graph 1, both B and C cause E (adapted from (*3*)).

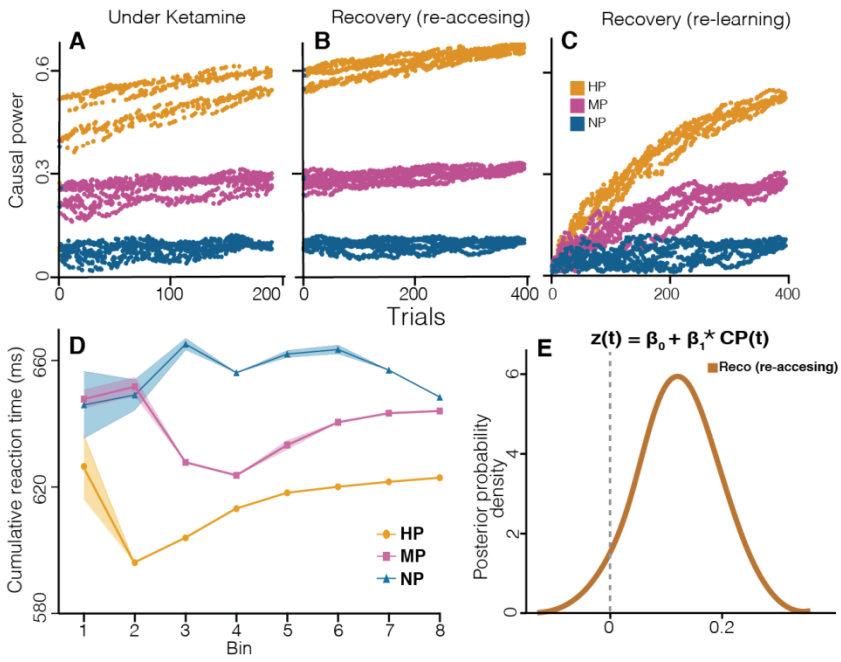

Fig. S2. Ketamine blocked access to predictive information. (A) Population causal power values for subjects under ketamine. (B) Population causal power values after recovery when subjects regained access to predictive information. (C) Population causal power values for hypothetical condition where subjects re-learned predictive information. (D) Binned population cumulative average RT (±SE) during first 200 trials after recovery from ketamine (25 trials per bin) for highly predictive (HP), moderately predictive (MP), and not-match predictive (NP) sounds. (E) Posterior probability density of β_1 for first 200 trials when subjects re-accessed predictive information (P=0.03).

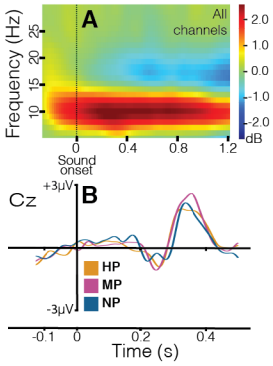

Fig. S3. Early sensory processing differences cannot account for prediction strength as all three sounds generated similar auditory ERPs. (A) Time-frequency plot of power averaged over all electrodes and all trials. Power calculated in 0.55 s sliding windows, with window at 0 s representing interval -0.275 s to +0.275 s. Plot aligned to sound onset. (B) Auditory ERPs at Cz electrode for highly predictive (HP), moderately predictive (MP) and not-match predictive (NP) sounds. Linear contrast of N200 for HP, MP and NP not significant (p=0.15).

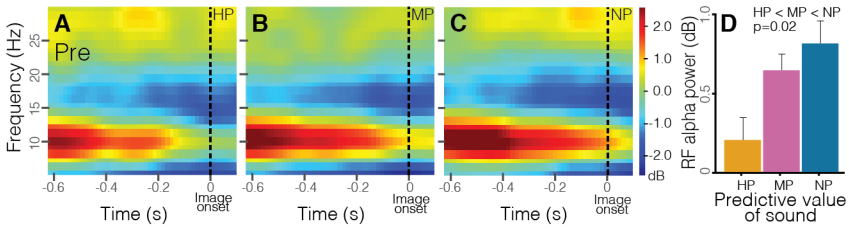

Fig. S4. Stronger predictions correlated with lower delay period alpha power at right frontal electrode cluster. Time-frequency decomposition of right frontal electrode cluster (RF) before drug administration (Pre) for HP (A), MP (B) and NP (C) sounds. Power calculated in 0.55 s sliding windows, with window at 0 s representing interval -0.275 s to +0.275 s. Plots aligned to image onset. (D) Population average RF alpha power (+SE) during delay period.

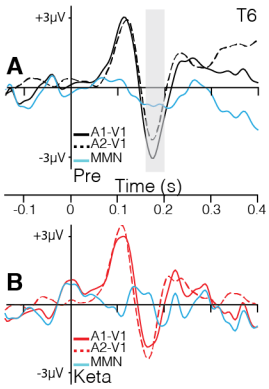

Fig. S5. Mismatch negativity disappears under ketamine. Population average visual ERP to V1 image at electrode 178 (T6) (A) before (Pre) and (B) under ketamine (Keta). Gray area in (A) highlights mismatch negativity (MMN) significantly less than zero (MMN<0; P=0.001). MMN in (B) was not significantly different from zero (MMN<0; P=0.18). P values FDR corrected.

	Effect Present (e^+) *Effect Present (V)*	Effect Absent (e^-) *Effect Absent*
Cause Present (c^+) *Cause Present (A)*	$N(c^+,e^+)$ *A1-V1*	$N(c^+,e^-)$ *A1-V2, A1-V3*
Cause Absent (c^-) *Cause Absent*	$N(c^-,e^+)$ *A2-V1, A3-V1*	$N(c^-,e^-)$ *A2-V2, A2-V3, A3-V2, A3-V3*

Table S1.

2 X 2 contingency table for each sound. Generic contingency table for all sounds in black. $N(c^+,e^+)$ represents the number of trials in which the effect occurs in the presence of the cause, $N(c^-,e^+)$ represents the number of trials in which the effect occurs in the absence of the cause, $N(c^+,e^-)$ represents the number of trials in which the cause occurs but not the effect, and $N(c^-,e^-)$ represents the number of trials in which the cause and effect are absent. 'Cause' used in statistical sense. In green, example contingency table for sound A1 where N(c+,e+) are the number of trials V1 follows A1 (A1-V1); whereas N(c-,e+) would be the number of trials V1 follows A2 or A3 (A2-V1 or A3-V1) and so on.