The Remainder Signature

A matched-data-quality test of whether the residual prediction error in conscious perception is real structure or a measurement artifact

Finn Petrov, Researcher in Consciousness science

International Academy for Consciousness Studies

Researched and prepared with the Academy. Working draft 1.0

Abstract

Predictive-processing accounts hold that cortex perpetually predicts its own sensory input and that conscious perception is bound up with the handling of prediction error. A recurring and rarely tested claim rides along with these accounts: that even after a well-fit generative model has explained what it can, a residual error signal survives, and that this residue is a real feature of conscious processing rather than a byproduct of noisy recording. The residue is easy to report and hard to interpret, because the two things that could produce it, genuine unmodeled structure in the stimulus and low measurement quality in the instrument, are almost always confounded in published designs. Richer stimuli tend to be recorded under different conditions than sparse ones, and prediction-error amplitude is known to scale with data quality. This paper proposes a decisive test that breaks the confound. We describe a matched-data-quality EEG and MEG protocol that holds recording quality, trial count, and signal-to-noise ratio (SNR) fixed by construction while varying only the structure of the stimulus across three regimes: predictable, structured-but-unpredictable, and noise. The prediction is that a residual prediction-error signal tracks stimulus structure, a "remainder signature," and not SNR. We state an explicit kill condition. If, under matched data quality, the residual always tracks SNR and never structure, the claim that the residue is a real feature of conscious processing does not survive. We tag every claim by epistemic status and firewall the neural result from any claim about phenomenal experience.

1. The residue that nobody has isolated

Predictive processing has become the default grammar for talking about perception in cortex. In its hierarchical-inference form, higher levels send predictions downward and lower levels return the mismatch, the prediction error, upward; learning and perception are the ongoing minimization of that error [Rao and Ballard 1999; Friston 2010; Clark 2013]. The framework is productive precisely because it makes error a measurable quantity. The mismatch negativity (MMN), the auditory event-related response to a violated regularity, is the workhorse here, and hierarchical predictive-coding models have been fit to it in some detail [Naatanen et al. 2007; Wacongne et al. 2011].

A quieter claim travels inside these accounts. It is the claim that error minimization never reaches zero, that after the model has done its explaining, something is left over, and that this leftover is not merely the tail of a noise distribution but a real, structured feature of how a conscious system processes the world. The claim is attractive. It gives the theory something to point at when it wants to say that perception is never a closed book, that the world always exceeds the model. It is also, as stated, nearly untestable, because the residue has never been isolated from the most boring possible explanation: that you were recording badly.

(FACT) Prediction-error amplitude, as measured by scalp EEG and MEG, scales with recording quality. Trial count, electrode impedance, subject movement, and background physiological noise all move the estimated error. (FACT) Stimulus richness and recording condition are correlated across the published literature, because richer paradigms often run longer, fatigue subjects differently, and are analyzed with different trial yields. (HYPOTHESIS) Therefore a large fraction of what has been reported as "residual" or "irreducible" prediction error may be data-quality variance wearing the costume of structure. This paper is about how to find out.

2. What the Closure lens contributes, and where it stops

We use one framing device and then set it down. In the Closure account, a system is openness (M) that settles into a definite content (c) under rules that keep it itself, a closure (Cl), leaving a remainder (R) that the settling did not absorb. Applied here, the map is exact and modest: the generative model is the closure, the predicted percept is the content, and the residual prediction error is a candidate remainder R.

The whole scientific question is what kind of thing R is. A remainder can be like weather, structured, lawful, carrying information about the world that the current model has not yet folded in; or it can be like a coin, a fair sample of noise that carries no structure at all and averages to nothing. The Closure lens does not decide this. It only sharpens the question and warns against a specific error: mistaking the fingerprint of your instrument for the fingerprint of the world. The rest of the paper is method, not metaphor. (WAGER) We bet R is weather. The point of the design below is that the bet can lose.

3. Why existing paradigms cannot answer the question

The local-global paradigm is the closest existing tool, and its limits show exactly what a new design must fix. Bekinschtein and colleagues separated two kinds of auditory regularity violation: a local violation, a deviant tone within a short sequence, and a global violation, a rare sequence type across seconds [Bekinschtein et al. 2009]. Local violations produced an early auditory response that did not depend on attention or awareness; global violations produced a late, distributed response present only when subjects were attentive and aware. Wacongne and colleagues extended this to a full hierarchy of predictions and prediction errors and showed the two levels interact: the local mismatch shrinks when the deviant becomes globally predictable [Wacongne et al. 2011].

These are strong results and we lean on them. But they were not built to separate structure from data quality, and they do not. The local and global conditions differ in more than the structure of the regularity; they differ in stimulus statistics, trial yield, and the attentional and arousal state of the subject, all of which

move measured error independently of any remainder. Attention itself, in the predictive-processing reading, is precision, the inverse-variance weighting of prediction error, so an attention difference is a data-quality difference in the model's own terms [Feldman and Friston 2010]. When a late response appears for the aware, structured condition, we cannot currently say whether we are seeing structure in R or better effective SNR in the aware state. (DISPUTED) The field routinely reads such effects as evidence for structured, consciousness-linked error signals. That reading is premature until data quality is held fixed by construction rather than by argument.

4. The core move: match data quality, vary only structure

The design rests on one discipline. Recording quality, trial count, and SNR are fixed across conditions by construction; the only thing that varies is the structure of the stimulus. Everything else in the protocol exists to make that sentence literally true.

We define three stimulus regimes, matched on every low-level physical statistic we can match (spectral power, envelope statistics, mean intensity, duration, event rate):

1. Predictable (P): a stimulus stream with learnable regularity, the kind a generative model can lock onto. A well-fit model should drive residual error toward its floor.

2. Structured-but-unpredictable (S): a stream with genuine higher-order structure that the model cannot reduce to a short-horizon prediction within the trial. This is the regime where a real remainder should be largest, because there is lawful structure that the closure fails to absorb.

3. Noise (N): a stream matched on all low-level statistics but carrying no learnable structure at any order. A fair coin.

The residual of interest is the prediction-error signal that survives after fitting each subject's generative model to their own data, condition by condition, using a hierarchical predictive-coding model of the class already fit to MMN data [Wacongne et al. 2011]. We are not measuring raw evoked amplitude. We are measuring what is left after the best available model has predicted as much as it can, which is the operational definition of R.

5. Holding SNR fixed by construction, not by hope

The claim that data quality is matched has to be earned mechanically. Five commitments do that work.

First, closed-loop trial equalization. Conditions are not run for a fixed duration and then compared; they are run until each contributes an identical number of artifact-clean trials, verified online. Trial count is the single largest driver of error-estimate reliability, and it is equalized by construction rather than by post-hoc matching.

Second, an injected SNR ceiling. Into every condition, including noise, we inject a common, physically identical calibration event (a fixed probe embedded at a fixed rate). The cortical and instrumental response to that probe is a per-subject, per-condition SNR meter. If the probe response differs across P, S, and N, data quality was not matched and the block is rejected and rerun. The probe is the load-bearing part of the design: it converts "we believe SNR was similar" into a measured, falsifiable equality.

Third, interleaving at the block level. P, S, and N blocks are interleaved within session so that drift in impedance, arousal, and fatigue is distributed evenly rather than confounded with condition. Order is counterbalanced across subjects.

Fourth, precision (attention) control. Because attention is precision in this framework and precision changes measured error [Feldman and Friston 2010], attentional state is held fixed with an orthogonal task that is identical across P, S, and N and does not itself depend on the manipulated structure. We monitor task performance as a covariate and discard blocks where it diverges across conditions.

Fifth, simultaneous EEG and MEG with a shared forward model. MEG and EEG have partly independent noise sources; a remainder that is real structure should appear in both under a shared source model, while an artifact of one instrument's noise floor need not. Convergence across modalities is a second, independent guard on the data-quality claim.

(HYPOTHESIS) With these five in place, the residual becomes a clean function of one variable, stimulus structure, because the others have been pinned. (WAGER) We expect the pinning to hold well enough that the probe response is statistically indistinguishable across conditions; if it is not, the experiment has failed on its own terms and reports nothing about R.

6. Prediction and kill condition

The test reduces to a single ordering of the residual across the three regimes, under verified-matched data quality.

> **Prediction.** Under matched data quality (identical artifact-clean trial counts, and an injected calibration probe whose cortical response is statistically indistinguishable across conditions), the residual prediction-error signal that survives per-subject model fitting will be ordered by stimulus structure and not by SNR: largest for structured-but-unpredictable (S), smallest for predictable (P), with noise (N) at or below the P floor. This structure-tracking residual, replicable across EEG and MEG under a shared source model, is the remainder signature. Concretely, residual(S) > residual(P) and residual(S) > residual(N), while the calibration-probe SNR is equal across P, S, and N.
>
> **Kill.** If, under verified-matched data quality, the residual does not track structure, that is, if residual(S), residual(P), and residual(N) are ordered by (or predicted by) the calibration-probe SNR and show no structure-dependent term once SNR is regressed out, or if the residual is statistically flat across P, S, and N, then there is no remainder signature. The claim that the residual prediction error is a real feature of conscious processing, rather than a data-quality artifact, does not survive. In particular, if residual(N) is greater than or equal to residual(S) at matched SNR, the "irreducible residue" reading is dead: the leftover is a coin, not weather.

Two features make this a real risk rather than a rigged win. The noise condition N is matched on every low-level statistic, so a residual that merely tracks physical richness would show up in N and kill the structure reading. And the calibration probe gives SNR its own measured channel, so "the residual tracks structure" can be tested with SNR partialed out, which is exactly the test that existing paradigms cannot run.

7. What a positive result would and would not show

Suppose the prediction holds: a structure-tracking residual, replicable across EEG and MEG, with SNR pinned. Firewall first. This result would show that a residual prediction-error signal carries information about stimulus structure that the fitted generative model did not absorb, under conditions where measurement quality cannot explain it. That is a claim about a neural signal and its informational content. It is not a claim about phenomenal experience.

(FACT, if obtained) The residual would be a structured neural quantity, not a noise artifact. (HYPOTHESIS) It would be a candidate correlate of the open, unclosed part of perceptual processing, the R of the Closure map. (Not licensed) It would not show that this residual is what it is like to perceive, that it is necessary or sufficient for consciousness, or that a system exhibiting it is conscious. Prediction-error signals occur in states and preparations where awareness is absent; the early local-mismatch response is the standard example, present without attention or report [Bekinschtein et al. 2009]. A remainder signature would inherit that caution. The honest statement of a positive result is narrow: a residual error signal tracks structure and not measurement quality. Whether that residual is a mark of conscious processing specifically is a further question, to be tested by crossing the paradigm with an awareness manipulation, not assumed.

This narrowness is a feature. The literature's habit of reading structured late responses as consciousness signatures is exactly the overreach this design is built to avoid. We report the signal's job description, not its inner life.

8. Limits, competing explanations, and the wager

Several things could go wrong or be misread, and naming them is part of the method.

Model dependence. The residual is defined relative to a fitted generative model. A different model class could relabel structure as prediction, shrinking R in S, or fail to fit P, inflating R there. (HYPOTHESIS) The result is only interpretable within a stated model family; we pre-register the model class and report residuals under at least two families to check that the structure ordering is not an artifact of one parameterization. If the remainder signature appears under one model and vanishes under another, that is informative and must be reported, not hidden.

Residual attention leakage. If the structured condition covertly recruits more precision despite the orthogonal task, some of residual(S) could be precision, not remainder [Feldman and Friston 2010]. The task-performance covariate and the probe-SNR meter are the guards; if they move with condition, the block is void. We would rather discard data than launder an attention effect into a remainder.

Structure that is secretly predictable. The S regime must contain structure the model genuinely cannot reduce within the trial. If subjects learn it across the session, S drifts toward P and the effect fades over time. We monitor the residual as a function of session time; a within-session decline in residual(S) toward the P floor is itself evidence about learnability and should be reported as such.

Generality. Auditory streams are the natural first instrument because the MMN and local-global tools are mature there [Naatanen et al. 2007; Bekinschtein et al. 2009]. Whether a remainder signature generalizes

to vision or to interoception is open. Interoceptive inference is a particularly interesting extension, because the body is a source of structured, hard-to-model input and interoceptive prediction error has been proposed as central to the feeling self [Seth 2013]; but that is a later paper, and importing it now would be exactly the overclaim we are trying to kill.

The wager. (WAGER) We are betting that when the confound between structure and data quality is finally broken by construction, a structure-tracking residual will remain. We could be wrong. If the residual collapses onto SNR the moment data quality is pinned, then a good deal of talk about irreducible residues in conscious perception was talk about noise, and the honest move is to say so. The design is built so that outcome is reportable and clean. A theory of perception that cannot lose this bet was never saying anything measurable; the value of the remainder signature is that it can fail.

9. Conclusion

Predictive processing gave us a measurable currency, prediction error, and with it a tempting story about a residue that conscious systems never fully discharge. The story has floated for want of a test that separates real structure from bad recording. We have proposed that test. Hold data quality fixed by construction, with equalized trial counts, an injected SNR probe, interleaving, a precision control, and simultaneous EEG and MEG; then vary only stimulus structure across predictable, structured-but-unpredictable, and noise regimes; then ask whether the model-residual tracks structure or SNR. If it tracks structure, there is a remainder signature, a modest, firewalled fact about a neural signal, and a first foothold for asking whether that signal is tied to awareness. If it tracks SNR, the irreducible residue was a coin all along. Either way, a claim that has lived on plausibility gets a chance to die on data. That exchange, plausibility for a kill condition, is the whole point.

References

Bekinschtein, T. A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., and Naccache, L. (2009). Neural signature of the conscious processing of auditory regularities. Proceedings of the National Academy of Sciences, 106(5), 1672-1677.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3), 181-204.

Feldman, H., and Friston, K. J. (2010). Attention, uncertainty, and free-energy. Frontiers in Human Neuroscience, 4, 215.

Friston, K. (2010). The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11(2), 127-138.

Naatanen, R., Paavilainen, P., Rinne, T., and Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. Clinical Neurophysiology, 118(12), 2544-2590.

Rao, R. P. N., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 2(1), 79-87.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. Trends in Cognitive Sciences, 17(11), 565-573.

Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L., and Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. Proceedings of the National Academy of Sciences, 108(51), 20754-20759.