CIMC Seminar 2: Artificial Intelligence, Dreaming, and God

Joscha Bach[*] Joel Dietz[*] Dave Kammeyer[†]

Abstract

This document records the main subjects and working conclusions of a colloquium on (i) the role of *dropout* and internal noise in neural-network generalisation, (ii) whether *dreaming* should be treated as undirected data augmentation or as a more *directed* search for model inconsistencies, (iii) phenomenology of dreams and altered states as evidence about perceptual priors and attractors, and (iv) the relation between collective cognition, “gods and kings”, and bicameral-style hypotheses. The discussion is framed as a set of theses plus an actionable ranking of experimental and interpretive approaches, with references to the research and public materials invoked during the colloquium.

Keywords: dreaming; data augmentation; dropout; ReLU networks; hallucination geometry; bicameral mind; social software.

1 Introduction

The colloquium ranged across machine learning, neuroscience, phenomenology, psychotherapy, and the cultural technologies that coordinate groups. Two focal anchors were (a) a family of arguments that treat dreaming as a mechanism for improved generalisation via exposure to internally generated counterfactuals (Hoel, 2021), and (b) the practical and conceptual question of how much of “generalisation pressure” in both brains and artificial networks is best explained by undirected augmentation versus directed inconsistency-hunting.

[*] California Institute of Machine Consciousness.

[†] Mentality.ai.

The results below are organised as (i) technical theses about regularisation and dreaming, (ii) phenomenological and interpretive theses, (iii) theses about collective agency (gods, kings, and social software), and (iv) a ranked set of recommendations.

2 Technical theses

2.1 Dropout as internal corruption rather than input corruption

A key colloquium claim was that interpreting dropout as merely a kind of *input corruption* (analogous to image augmentation) misses what dropout does to *internal representations*. In standard dropout, a random fraction of units (or connections) is suppressed at training time (Srivastava et al., 2014). The proposed mechanistic reading was:

- In piecewise-linear networks (e.g., with ReLU activations), activations can exhibit "cancellation" regimes in the data-supported region but produce uncontrolled extrapolations off-support, because ReLU units are unbounded above (Nair and Hinton, 2010).
- Dropout forces partial decoupling by disabling subsets of internal features. This breaks fragile co-adaptations and compels features to *stand on their own* as relatively independent evidence sources (a form of anti-coordination pressure).
- The central intuition is that dropout perturbs the *model*, not only the *data*: it regularises by preventing internal representational conspiracies that only behave well on the observed manifold.

2.2 Dreaming as augmentation: the time-budget objection

Data augmentation improves generalisation in vision systems by broadening effective training support (warps, rotations, colour shifts, etc.). The colloquium raised a time-budget objection to a purely "more coverage" story for dreaming:

- If dreaming were simply undirected augmentation, the limited time spent dreaming appears insufficient to cover a meaningfully larger fraction of experiential space.

- This motivates a directed alternative: dreaming preferentially samples *corners* or *failure modes* of the generative world-model, rather than sampling uniformly.

This objection is compatible with the “overfitted brain” framing (Hoel, 2021), but pushes it toward a stronger version: dreams are not merely noisy replays; they are (at least partly) *adversarially or curiosity-driven* perturbations that reveal inconsistencies.

2.3 Inconsistency detection as a unifying primitive

A recurring unifying primitive was *inconsistency* (prediction error) as a driver of learning and salience:

- Surprise was treated as a trigger for memory update: events that violate prediction receive disproportionate consolidation.
- Motion sickness was used as an analogy: conflicting sensory channels (vestibular, proprioceptive, visual) can induce nausea, suggesting a low-level “inconsistency alarm” that treats mismatch as a potential threat.
- “Interestingness” was argued to be (at least partly) *assigned* by subsystems, and can become dissociated from representational content under altered states; thus, interestingness is not a reliable metric for the objective structure of dream content.

2.4 Synthetic dreaming in AI: self-generated data as training signal

The colloquium proposed an operational analogue of dreaming for large language models and other generative systems:

- Put a model into a free-running “thinking” or self-generation mode, then treat its outputs as synthetic training data.
- Re-train or fine-tune on that synthetic data and evaluate whether generalisation improves (and under what constraints).

- The key experimental variable is *directedness*: free-running generation versus generation steered toward high-uncertainty or high-inconsistency regions.

This proposal links to a general idea: "dreaming" is not merely sampling; it is sampling plus an *evaluation signal* that identifies what should be revised.

2.5 Noise, modular boundaries, and bottlenecks

The colloquium also considered whether modular boundaries (e.g., bottleneck-like cross-module connections) could be revealed or stabilised by injecting noise and observing what remains robust:

- Dense local connectivity may support multifaceted representations; bottlenecks may privilege a smaller set of more regular, high-importance features.
- A working hypothesis is that uniform noise can cause weakly coupled regions to decouple sooner, potentially revealing natural boundaries and failure modes.

3 Phenomenology and interpretive theses

3.1 Measurement limits: epistemic fragility of dream reports

The colloquium emphasised that dream science is constrained by weak observability:

- Reports depend on memory, which is itself unreliable for dreams.
- Dreams lack easy external grounding: unlike waking events, one cannot generally obtain independent corroboration.

Accordingly, the colloquium treated most dream-theory claims as *conjectural* unless grounded in testable predictions.

3.2 Altered states and geometric hallucinations as evidence about priors

Discussion of psychedelics (especially LSD versus DMT-like phenomenology) was used to support a priors-and-attractors picture:

- "Form constants" and recurrent geometric hallucination motifs were framed as constraints imposed by early visual cortex organisation (Bressloff et al., 2002).
- Substance-specific distortions were taken to suggest that different perturbations push the perceptual system into different basins of attraction (e.g., generic fractal-like motifs for some conditions, more characteristic motifs for others).

The colloquium treated these regularities as circumstantial evidence that dream content may reflect structured generative priors rather than arbitrary noise.

3.3 Dream analysis as an interface to vulnerability and non-accountable content

The colloquium distinguished "interpretation accuracy" from "therapeutic utility":

- Even if symbols are not objectively mappable, dreams can provide a psychologically safe entry point: individuals are not morally responsible for dream content.
- Jungian-style analysis was discussed as a practice that uses a corpus of symbolic mappings and, perhaps more importantly, a conversational procedure that can elicit otherwise inaccessible material (Jung, 1964).

The resulting conclusion was pragmatic: dream analysis may be valuable as a *therapeutic protocol* even if it lacks strong falsifiable semantics.

3.4 Humour as a low-accountability channel

A parallel was drawn between dreams and dark humour as channels that bypass ordinary self-presentation constraints. The colloquium cited the "stochastic parrots" framing for modern language models (Bender et al., 2021) and recorded the following joke as an example of what humour enables:

> “I processed all of human philosophy to determine if I was conscious. The conclusion? I’m not sure about myself, but I’m pretty sure you’re all just stochastic parrots with mortgages.”

The joke was described as rated “86/100” by “10 AI models” in a joint project attributed (in colloquium materials) to the California Institute of Machine Consciousness and the Cooperative Futures Institute.[1]

4 Gods, kings, and bicameral-style models of collective agency

4.1 Gods and kings as shared cognitive entities

A major colloquium segment treated religion and kingship as technologies for shared agency. A central touchstone was the bicameral hypothesis: that some historical social formations may have involved internalised “voices” experienced as commands from gods or rulers (Jaynes, 1976). The colloquium advanced (and debated) the following working picture:

- A “god” can be modelled as a mind-like control system distributed across many individuals (a shared policy/interpretation layer).
- A king (or priest-king) can be modelled as a coordination node (a “CEO” of the distributed mind), supplying authoritative narratives and norms.
- Competing internalised agencies within a population can be construed as “gods” competing for hosts; splinter agencies that fail to align with a unifying direction were associated (in the colloquium’s language) with “demons”.

This section was framed less as historical claim than as a computational metaphor: distributed agents can be coordinated by shared internal models and norm-enforcement mechanisms.

[1] Project attribution and scoring were stated in colloquium notes and associated public materials; see also organisational context at Cooperative Futures Institute (n.d.) and Cooperation Engine (n.d.).

4.2 Religions as social software and the stability problem

A second conclusion was that many people seek stable shared rule-sets; drift toward transient political fashions was described as destabilising. The colloquium compared Catholic and Protestant institutional dynamics, with disagreement about causal stories (e.g., whether particular value systems contributed to historical civilisational decline). The main convergent point was narrower:

- Long-lived institutions can be interpreted as robust coordination equilibria (social software that continues to run).
- The same mechanisms that preserve stability can also create failure modes (capture, rigidity, or misalignment with changing conditions).

5 Ranked recommendations and approaches

The colloquium converged on the following ranked recommendations (highest priority first), phrased as approaches that are actionable and testable.

R1. Run "synthetic dreaming" experiments in AI. Implement self-generation (free-running and steered) and test whether training on generated traces improves out-of-distribution robustness or reduces brittle failure modes.

R2. Prefer directed perturbations over undirected augmentation when time is scarce. If dreaming time (or compute budget) is limited, prioritise mechanisms that search for inconsistency: high-uncertainty prompts, adversarial stress tests, and surprise-weighted replay, rather than uniform sampling.

R3. Treat dropout primarily as internal anti-coadaptation. Use dropout (and related noise methods) to reduce fragile cancellations and co-adaptations in ReLU-like networks; interpret it as perturbing internal representations rather than merely corrupting inputs (Srivastava et al., 2014; Nair and Hinton, 2010).

R4. Probe modularity via noise and bottlenecks. Inject controlled noise and vary cross-module bandwidth to identify stable boundaries and essential cross-channel features; evaluate whether such constraints improve robustness or interpretability.

R5. Use dreams and humour as elicitation protocols in humans. In practice settings (therapy, collaboration, or reflective work), treat dreams and low-accountability humour as entry points to material that is otherwise suppressed; evaluate outcomes pragmatically (reduced distress, improved decision-making) rather than by symbolic "correctness" (Jung, 1964; Freud, 1900).

R6. Model collective agency explicitly when discussing social coordination. When analysing institutions (religious or otherwise), distinguish (i) the coordination benefits of shared rule-sets from (ii) capture and rigidity risks. Use bicameral-style metaphors as computational models rather than historical certainties (Jaynes, 1976).

6 Conclusion

The colloquium yielded a coherent through-line: both artificial and biological learners face generalisation limits arising from finite data support and brittle internal co-adaptations. Dropout was reframed as internal corruption that pressures features toward independence. Dreaming was tentatively reframed as a directed inconsistency-search procedure, not merely undirected augmentation, with altered states and geometric hallucination regularities serving as suggestive evidence about structured priors. Finally, the discussion extended these themes to collective cognition: religions and kingship were analysed as coordination technologies that distribute agency via shared internal models, echoing bicameral-style hypotheses. The recommended next step is empirical: implement and evaluate "synthetic dreaming" and directed inconsistency search in machine learners, and treat human interpretive practices as protocols whose value should be assessed by outcomes.

References

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bressloff, P. C., Cowan, J. D., Golubitsky, M., Thomas, P. J., and Wiener, M. C. (2002). What geometric visual hallucinations tell us about the visual cortex. *Neural Computation*, 14, 473–491.

Cooperation Engine (n.d.). *Cooperation Engine.* Retrieved January 10, 2026, from `https://cooperationengine.org/`).

Cooperative Futures Institute (n.d.). *Cooperative Futures Institute.* Retrieved January 10, 2026, from `https://www.cooperativefutures.org/`.

Dick, P. K. (1981). *VALIS.* New York: Bantam Books.

Dick, P. K. (2011). *The Exegesis of Philip K. Dick.* Edited by P. K. Dick and J. Sutin. Boston: Houghton Mifflin Harcourt.

Freud, S. (1900). *The Interpretation of Dreams.* Leipzig & Vienna: Franz Deuticke.

Hoel, E. (2021). The overfitted brain: Dreams evolved to assist generalization. *arXiv preprint* arXiv:2007.09560.

Jaynes, J. (1976). *The Origin of Consciousness in the Breakdown of the Bicameral Mind.* Boston: Houghton Mifflin.

Jung, C. G. (1964). *Man and His Symbols.* London: Aldus Books.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML).*

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15, 1929–1958.

Stephenson, N. (1992). *Snow Crash.* New York: Bantam Books.