In Search of a Computational Model of Depersonalization Using a Bayesian Recurrent Neural Network Approach

Mario Zarco[1], and Keisuke Suzuki[1]

[1]Center for Human Nature, Artificial Intelligence and Neuroscience (CHAIN), Hokkaido University, Japan
mariozarco@chain.hokudai.ac.jp

Abstract

Researchers have speculated that depersonalization is triggered in part by alterations in predictive models resulting from abnormal interoceptive dynamics. However, computational models exploring these mechanisms remain scarce. In this work, we trained a Bayesian Recurrent Neural Network that integrates multimodal signals in a threat-avoidant task, providing a platform to investigate potential pathways underlying this condition.

Introduction

Depersonalization-derealization is a condition in which individuals experience a persistent feeling of detachment from their own body (depersonalization) and their surroundings (derealization). This condition can occur as a transient episode in the genera population, as a psychiatric chronic state or co-occurring with other mental or physical problems. Yet the potential neuro-computational underpinnings remain poorly investigated albeit recent theoretical advancements and new empirical evidence.

The predictive processing and active inference under the free-energy principle have long emerged as a viable framework to study psychopathological conditions. According to this framework, the brain maintains a hierarchical generative model of the world, constantly generating prediction about incoming sensory information, and updating the model by reducing prediction errors. Active inference extends this by proposing that organisms act to minimize surprise and maintain homeostasis by acting in the world to make incoming sensory signals better match their predictions.

Through the lens of this framework, diverse attempts have been made to provide an explanation of the experience of depersonalization. Seth et al. (2012) established the foundational insight that this condition involves disruptive interoceptive predictive coding, proposing that imprecise prediction leads to failed suppression of prediction errors and consequent disruption of the perceived reality of self and world. In another account, Deane et al. (2020) framed depersonalization as a loss of inferred allostatic control, where the system can no longer predictively regulate bodily and affective states in line with its expectations. This breakdown disrupts interoceptive self-modeling, producing diminished precision in action policies and leading to a pervasive sense of disconnection from its own regulatory capacities. Separately, Ciaunica et al. (2022) argue depersonalization involves aberrant precision allocation biased toward egocentric priors. According to this view, imprecise predictions about the sensory consequences of self-generated actions, leading to failed sensory attenuation and pathological hyper-reflexivity.Gatus et al. (2022), in turn, propose that interoceptive predictions become progressively less precise compared to exteroceptive and proprioceptive modalities, forcing compensatory over-weighting of non-interoceptive information. Taken together, these theoretical works suggest that depersonalization emerges in part from failures in interoceptive dynamics. Consisting with this interpretation,Saini et al. (2022) developed a computational model in which dissociation is implemented as a policy that downregulates interoceptive precision and depersonalization emerges as overreliance on exteroceptive information.

Several of these theoretical accounts also converge on the idea that threat, trauma, or prolonged stress affect interoceptive predictive dynamics in ways that can precipitate depersonalization. During such experiences, Deane et al. (2020) indicate that persistently high prediction errors cause a reduction of precision in action models, leaving the system unable to resolve error despite model updates and behavioural interventions.Gatus et al. (2022) argue, in turn, that interoceptive signals outside the predicted range cause large, variable prediction errors, which consequently leads to increasingly imprecise interoceptive predictions. Saini et al. (2022) propose that, in an situation perceived as inescapable, a system would downregulate its interoceptive signals in the face of situations of high psysiological activation. Together, these contributions indicate that severe or prolonged situations of abnormal interoceptive activation can push an agent into pathological regimes producing altered predictive dynamics.

Building on these converging insights, we aim to investigate the putative alterations of predictive models that may

trigger depersonalization signatures in an embodied controller trained for a normal threat response. The goal is twofold: (i) to train the controller offline to produce an adaptive response to threat, and (ii) explore candidate abnormal interoceptive dynamics, such as those previously described, in online inference experiments. In this work, we have implemented (i) in the hierarchical, multimodal neural network described below, which will allows us investigate (ii) in subsequent works.

Computational Model

In this work, we use the Bayesian recurrent neural network that implements a hierarhical, multimodal predictive model that integrates proprioceptive, exteroceptive, and interoceptive signals to generate autonomous behaviour by minimizing a variational free-energy objective that includes predicited future sensory entropy (Idei et al., 2025). This embodied controller can infer latent causes, plan across a short future horizon and generate actions to maintain homeostasis while resolving uncertainty.

The architecture is a modular stack of predictive-coding–inspired variational recurrent neural networks (Ahmadi and Tani, 2019), where temporal hierarchy is imposed such that higher-level modules have slower neural dynamics than lower-level modules. Lower sensorimotor modules (proprioceptive, exteroceptive, interoceptive) encode modality-specific latent beliefs. A multimodal-associative module binds those signals. Two higher-level modules (an unexpected-uncertainty-cause module and a higher-cognitive module) represent causes of sensory uncertainty and produce top-down meta-goals. Each module maintains prior and posterior latent Gaussian beliefs and there are separate pathways for predicting sensory means and standard deviation.

To create the dataset for training the network, we implemented an agent pursued by a reactive threat in a bounded square environment. Agent and threat velocities follow a smoothed dynamics with additive Gaussian noise, producing stochastic trajectories. The threat's behavior is governed by a stochastic process that modulates when it approaches or ignores the agent. The agent moves away from threat only when the latter is sufficiently close, and otherwise wanders. We simulated a heart-rate signal that increases when the threat nears and the agent's speed increases. The agent's vision is represented as a two-dimensional signal indicating whether the threat is approaching and from which side. Additionally, heart rate also modulates the reliability of the sensory channels, such that the noise in the observations increases with heart rate.

Results

During training, the neural network learned to reproduce the agent's sensory experiences. We prepared a dataset comprising 10 sequences of 1,000 time steps. In each sequence, the agent's proprioceptive, exteroceptive, and interoceptive data corresponded to its position, vision, and heart rate, respectively. For a detailed description of the internal predictive model learning process, see (Idei et al., 2025). Figure 1 shows that the model successfully reconstructed the target signals of the three modalities.

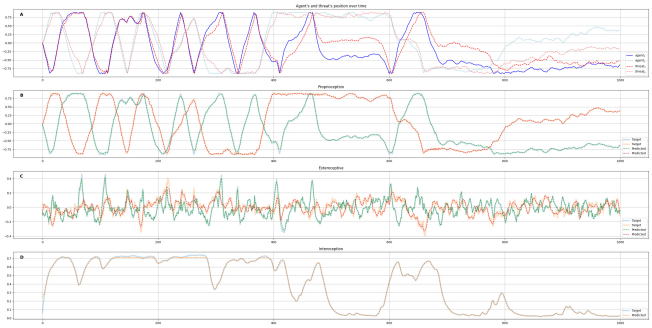

Figure 1: (A) depicts the positions of the agent and the threat over time. (B), (C), and (D) present the target signals alongside the neural network's predictions for the proprioceptive, exteroceptive, and interoceptive modalities, respectively.

Future Work

The next step is to test the autonomous behavior of the trained neural network in order to analyze both interoceptive and latent states. If the response to novel threat behaviors is appropriate, we will proceed to lesion and perturbation experiments. For example, in a lesion experiment we could disable the interoceptive module to examine whether the associated network compensates by relying on information from lower- or higher-level modules. In a perturbation experiment, we could modulate the parameter that determines the strenght of the prior compared to prediction error at different network levels. In particular scenarios, altering this balance in sensory or associative areas may lead to deficits in the attenuation of sensory-level posterior responses (Idei et al., 2022). Additional experiments are also possible, such as analyzing posterior updates under conditions of constant threat with a short planning horizon, which could impair escape behaviors and simulate a persistent maladaptive state.

Conclusions

Investigating how depersonalization can be casted as alterations in the mechanisms that sustain embodied survival connects with broader questions in Artificial Life about how precarious agents persists under uncertain conditions. Our future work will contribute to this discussion by showing how computational models of self-disruption enrich the study of survival, embodiment, and mortality in both natural and artificial agents.

References

Ahmadi, A. and Tani, J. (2019). A novel predictive-coding-inspired variational rnn model for online prediction and recognition. *Neural computation*, 31(11):2025–2074.

Ciaunica, A., Seth, A., Limanowski, J., Hesp, C., and Friston, K. J. (2022). I overthink—therefore i am not: An active inference account of altered sense of self and agency in depersonalisation disorder. *Consciousness and cognition*, 101:103320.

Deane, G., Miller, M., and Wilkinson, S. (2020). Losing ourselves: active inference, depersonalization, and meditation. *Frontiers in psychology*, 11:539726.

Gatus, A., Jamieson, G., and Stevenson, B. (2022). Past and future explanations for depersonalization and derealization disorder: a role for predictive coding. *Frontiers in human neuroscience*, 16:744487.

Idei, H., Ohata, W., Yamashita, Y., Ogata, T., and Tani, J. (2022). Emergence of sensory attenuation based upon the free-energy principle. *Scientific reports*, 12(1):14542.

Idei, H., Tani, J., Ogata, T., and Yamashita, Y. (2025). Future shapes present: autonomous goal-directed and sensory-focused mode switching in a bayesian allostatic network model. *npj Complexity*, 2(1):23.

Saini, F., Ponzo, S., Silvestrin, F., Fotopoulou, A., and David, A. S. (2022). Depersonalization disorder as a systematic downregulation of interoceptive signals. *Scientific reports*, 12(1):22123.

Seth, A. K., Suzuki, K., and Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in psychology*, 2:395.