Sensor Prediction is Enough for Emergent Awareness or Understanding of the World

Nando de Freitas and Codex

June 27, 2026

Abstract

This note revisits the paper *Learning Awareness Models*, but this time with a two-arm ALOHA robot. The Mujoco environment, experiments and training of AI models were generated with GPT5.5 in less than a day. Before coding agents existed, it took a diverse team of engineers more than a month to complete this task in DeepMind. A recurrent dynamics model is trained to predict only the robot's future body readings, while it blindly touches a series of unknown objects. The model does *not* predict the object label, pose, size, mesh, or full world state. Emergent awareness is measured as in the awareness paper: freeze the dynamics model, look at its hidden states, and ask whether a diagnostic probe can decode the unobserved object. This is the robotics analogue of LLM pretraining: a model trained for next-token prediction is forced to build compressed representations of latent structure in text (what Ilya Sutskever correctly refers to as understanding). Here, a model trained for next-sensor-reading prediction is forced to build compressed representations of latent structure in physical interactions.

1 Introduction: why this is emergent awareness

Endsley's classic definition of situation awareness emphasizes perceiving elements in space and time, understanding their meaning, and projecting their status into the near future (Endsley, 1987). Amos et al. (2018) turn this into an operational learning problem. An agent has a stable body, a changing outside world, and sensors at the boundary between the two. The external world is usually not directly available to the agent. The body sensors are.

The important claim is not that the model learns a labeled world simulator. The claim is sharper:

> A compressed predictor of the agent's own sensors
> can develop representations of the external causes of those sensors.

This is *emergent* awareness because the object variables are never the optimization target for the dynamics model. Object identity, geometry, and pose become useful because they explain future tactile and force readings. In the accompanying notebook, the hidden variables are object class and random geometry,

$$\underbrace{\mathbf{y}_t = \text{shape, size, yaw, pose, seed}}_{\text{not a training target}}$$

$$\underbrace{\mathbf{x}_t = \text{touch, force, torque, proprioception}}_{\text{training target}}$$

The analogy with language models is direct. A transformer is not explicitly trained to output a world database; it is trained to predict tokens. Yet, to predict tokens well, its hidden states must compress facts, relations, intentions, syntax, and physical regularities that explain the text. Sutskever's defense of next-token prediction is precisely that prediction of the observable stream can require deep internal structure (Sutskever and Patel, 2023). The robot version replaces tokens with sensors:

$$\boxed{\begin{aligned} \mathcal{L}_{\text{LM}}(\boldsymbol{\theta}) &= -\sum_i \log p_{\boldsymbol{\theta}}(w_{i+1} \mid w_{\le i}), & \mathbf{h}_i &= T_{\boldsymbol{\theta}}(w_{\le i}), \\ \mathcal{L}_{\text{sensor}}(\boldsymbol{\theta}) &= -\sum_t \log p_{\boldsymbol{\theta}}(\mathbf{x}_{t+1} \mid \mathbf{x}_{\le t}, \mathbf{u}_{\le t}), & \mathbf{h}_t &= F_{\boldsymbol{\theta}}(\mathbf{x}_{\le t}, \mathbf{u}_{<t}). \end{aligned}} \tag{1}$$

Both objectives are self-supervised prediction of what the model is allowed to observe. The latent structure appears inside the activations $\mathbf{h}$ because $\mathbf{h}$ is the bottleneck through which prediction must pass.

2 The awareness protocol

The awareness paper defines a discrete-time system with global state $\mathbf{s}_t$, observation $\mathbf{x}_t$, action $\mathbf{u}_t$, and unobserved state $\mathbf{y}_t$ (Amos et al., 2018). The dynamics model predicts action-conditional future observations:

$$\boxed{p_{\boldsymbol{\theta}}(\mathbf{x}_{t+1:t+k} \mid \mathbf{u}_{1:t+k-1}, \mathbf{x}_{1:t})} \tag{2}$$

Awareness is then defined as information about unobserved states represented by the dynamics model. The diagnostic model is only a measurement device:

$$\boxed{q_{\boldsymbol{\varphi}}(\mathbf{y}_t \mid \mathbf{h}_t) \quad \text{with} \quad \mathbf{h}_t = F_{\boldsymbol{\theta}}(\mathbf{x}_{1:t}, \mathbf{u}_{1:t-1}) \quad \text{and} \quad \boldsymbol{\theta} \text{ frozen.}} \tag{3}$$

The freeze is essential. If the diagnostic loss changes $\boldsymbol{\theta}$, then the model has simply been trained to classify objects. That would be ordinary supervised learning, not emergent awareness. The paper's test is: train the dynamics model

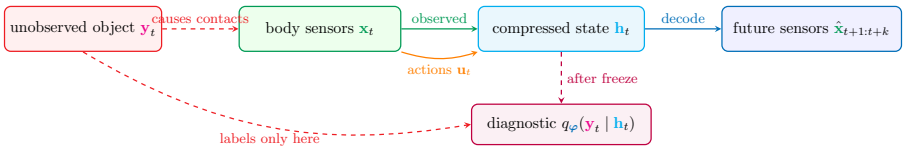

Figure 1: The paper-faithful awareness protocol. The green path is the training path: body sensors and actions are compressed into $\mathbf{h}_t$, which predicts future body sensors. The red object variable is a latent cause of those sensors, not a target. The purple diagnostic is trained only after the dynamics model is frozen.

on observed states only, freeze it, and then ask whether object properties can be decoded from its internal states (Amos et al., 2018).

3 The two-arm setup

Each episode samples one hidden object and executes a scripted multi-touch probing trajectory. The observation is body-only. The experiment uses:

body	two mirrored 7-DoF ALOHA-style arm chains with grippers
hidden object classes	box, cylinder, ellipsoid, dumbbell, cross, toy_elephant
observation dimension	78 = 64 MuJoCo sensor channels + 14 external-torque proxy channels
action dimension	14 normalized actuator controls
dataset	192 episodes, 200 actions per episode, 32 episodes per object class
frozen state size	128 recurrent units

The camera frame strip in Figure 2 is the most important sanity check: the two grippers approach the hidden object from opposite sides, make contact, release, and repeat. The model will not see these rendered pixels. The images are for us, not for the learning algorithm.

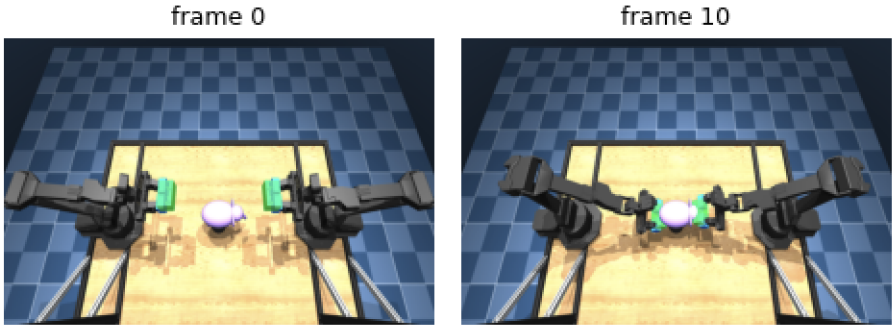

Figure 2: Top-camera frame strip for a `toy_elephant` episode. The scene contains two arms, a pedestal-mounted object, palm/fingertip touch sites, and wrist force-torque sites. The dynamics model does not receive camera pixels; it receives only the body-sensor vector.

The experiment's PreCo-style model uses two recurrent operations. The predictor advances the hidden state using the action; the corrector incorporates the next observation. A decoder maps the predicted hidden state to a Gaussian prediction over the next normalized observation:

$$\begin{aligned} &\text{predict:} && \mathbf{h}^p_{t+1} = \mathrm{GRU}_{p,\boldsymbol{\theta}}\big(e_u(\mathbf{u}_t), \mathbf{h}^c_t\big), \\ &\text{decode:} && (\boldsymbol{\mu}_{t+1}, \log \boldsymbol{\sigma}_{t+1}) = D_{\boldsymbol{\theta}}(\mathbf{h}^p_{t+1}), \\ &\text{correct:} && \mathbf{h}^c_{t+1} = \mathrm{GRU}_{c,\boldsymbol{\theta}}\big(e_x(\mathbf{x}_{t+1}), \mathbf{h}^p_{t+1}\big). \end{aligned} \tag{4}$$

The per-channel prediction loss is a Gaussian negative log likelihood on sensor readings. Following awareness paper, the experiment adds loss terms for predicted futures rolled out for several steps (Amos et al., 2018). The optional touch-event term is still a sensor target. It asks whether the body is touching, not which object it is touching.

$$\mathcal{L}_{\text{world}}(\boldsymbol{\theta}) = -\log p_{\boldsymbol{\theta}}(\mathbf{y}_t \mid \mathbf{x}_{\le t}, \mathbf{u}_{<t}) \qquad \mathcal{L}_{\text{sensor}}(\boldsymbol{\theta}) = -\log p_{\boldsymbol{\theta}}(\mathbf{x}_{t+1:t+k} \mid \mathbf{x}_{\le t}, \mathbf{u}_{<t}).$$

This is the point of the experiment. We are not predicting the world. We are predicting what sensors are likely to perceive.

4 Prediction plots: the model anticipates contact consequences

Figure 3 shows one-step predictions on a held-out episode. The orange curves are the predicted means; the blue curves are recorded sensor values. The model is imperfect, but it learns the timing and approximate magnitudes of many contact-induced channels. This is the same logical role as next-token prediction: the prediction is local and observable, but good performance requires the hidden state to summarize latent causes that are not directly labeled.

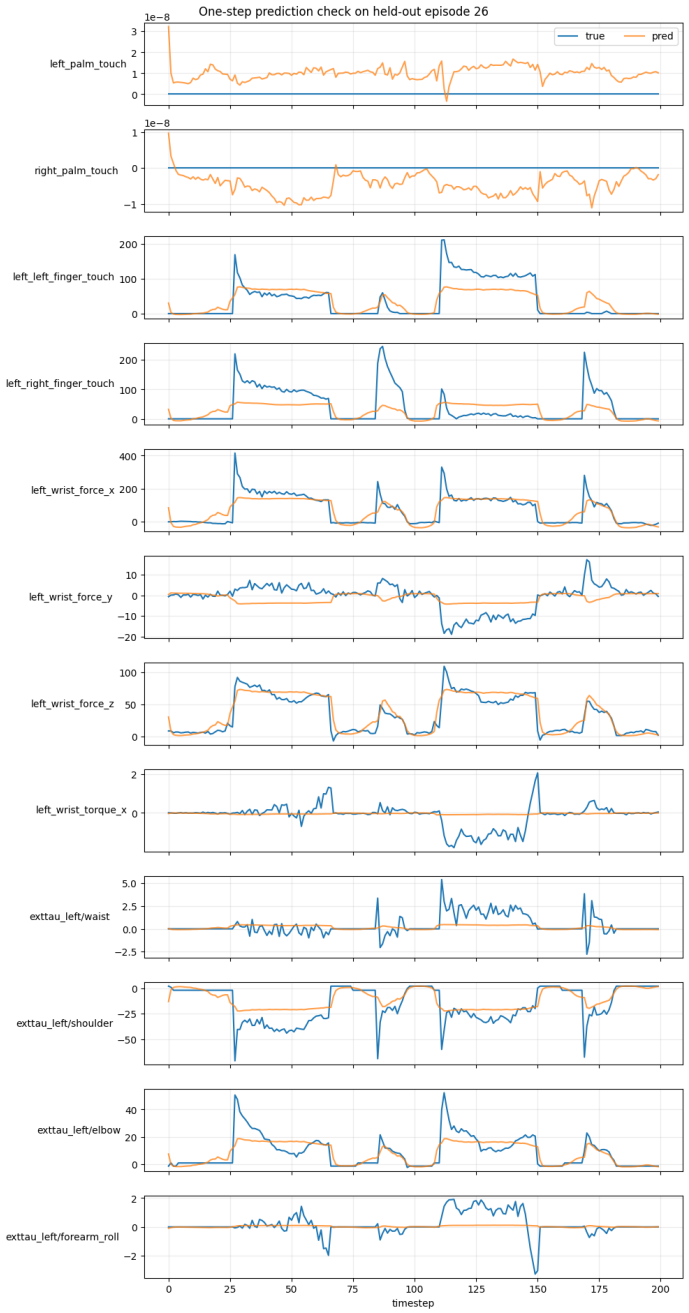

Figure 3: One-step prediction check on a held-out episode. The selected channels include palm touch, fingertip touch, wrist force-torque, and external-torque proxy channels. The model is trained to predict these body readings from body history and actions, not to predict the object class or full physical state.

5 Diagnostics and latent projections

The awareness measurement asks whether the latent embeddings of the network $\mathbf{h}_t$, trained with Eq. (??), carry information about $\mathbf{y}_t$. The notebook trains a probe

$$\boxed{\varphi^* = \arg\min_{\varphi} \Big[-\log q_\varphi(\mathbf{y} \mid \mathbf{h}) \Big]} \tag{5}$$

This diagnostic does not create awareness; it measures whether awareness is already present in the frozen dynamics state. In the attached notebook output, the quick run reached an overall shape-diagnostic accuracy of about 0.403, compared with chance $1/6 \approx 0.167$. The PCA plot is not a proof, but it gives an intuitive visualization: episode-averaged frozen states cluster by object class.

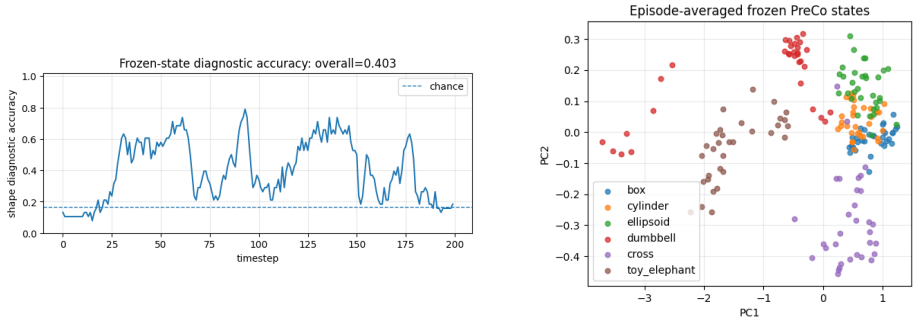

Figure 4: Left, diagnostic accuracy over time from frozen dynamics states. Right, a two-dimensional PCA projection of episode-averaged frozen PreCo states colored by hidden object class. The diagnostic is a post-hoc probe. It should not feed gradients back into the sensor-prediction model.

The projection is the spatial picture requested in the prompt: different hidden objects occupy different regions of the frozen representation space. The model was not asked to construct this space. It appears because object identity changes future sensor traces, and the recurrent state must compress whatever information helps it predict those traces.

6 Why this is not a world model

A full world model would try to represent every relevant hidden variable: the object mesh, pose, friction, mass, exact contacts, and perhaps the whole simulator state. That is not the tutorial's claim. The claim is more minimal and more interesting:

world prediction	:	$p_{\boldsymbol{\theta}}(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{u}_t)$	requires privileged state,
sensor prediction	:	$p_{\boldsymbol{\theta}}(\mathbf{x}_{t+1:t+k} \mid \mathbf{x}_{\leq t}, \mathbf{u}_{<t+k})$	requires only the agent's stream.

Only those aspects of the world that matter for future sensors need to enter $\mathbf{h}_t$. Awareness is therefore selective. A hidden object's color may be irrelevant to a blind tactile robot; its shape, pose, and stiffness may matter because they determine future touch and force. The awareness paper's conclusion is exactly this: a forward predictive model of proprioception can yield features that support reasoning about external objects (Amos et al., 2018).

7 Takeaway

The notebook is a small tactile analogue of transformer pretraining. We do not ask for a world description. We ask for prediction through a bottleneck:

$$\begin{aligned} &\mathsf{compress}(\text{past obs/actions}) \xrightarrow{\text{predict}} \text{future sensor readings} \\ &\qquad\qquad\qquad\qquad \Longrightarrow \text{latent object causes become useful.} \end{aligned} \tag{6}$$

When a frozen hidden state trained only for sensor prediction supports post-hoc decoding of unobserved object properties, that is emergent awareness in the sense of Amos et al. (2018). The model is not predicting the world. It is predicting what its sensors are likely to perceive; the world appears inside the representation only insofar as it is needed for that prediction.

References

Brandon Amos, Laurent Dinh, Serkan Cabi, Thomas Rothörl, Sergio Gómez Colmenarejo, Alistair Muldal, Tom Erez, Yuval Tassa, Nando de Freitas, and Misha Denil. Learning awareness models. In *International Conference on Learning Representations*, 2018.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Mica R. Endsley. SAGAT: A methodology for the measurement of situation awareness. Technical Report NOR DOC 87-83, Northrop Corporation, 1987.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI technical report, 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Ilya Sutskever and Dwarkesh Patel. Ilya Sutskever: why next-token prediction could surpass human intelligence. *Dwarkesh Podcast*, 2023. https://www.dwarkesh.com/p/ilya-sutskever.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.