

Praveen Suthaharan

Singapore · praveen.suthaharan@yale.edu · github.com/psuthaharan · psuthaharan.github.io

I build safe AI systems to support human learning. My current work is a multi-agent workflow that strengthens metacognition in adult learners, designed so that every decision it makes can be inspected and traced to the evidence. My academic training is in computational psychiatry and biostatistics, where I infer hidden states of the mind from brain activity and behavior to make sense of human mental health.

AI SAFETY, INTERPRETABILITY & EVALUATION

Mechanistic interpretability of Gemma-2-2B

Accepted, BlackboxNLP 2026 (ACL proceedings)

- Characterised how five dimensions of metacognitive scoring are encoded across a 26-layer residual stream using linear probing (ridge regression, 5-fold CV, length-residualised), logit-lens vocabulary coupling, and activation patching against a behavioural readout direction.
- Found a dissociation with direct bearing on evaluation validity: the most reliably encoded dimension (*Evaluation*, $r = 0.58$) showed near-zero behavioural coupling ($r = 0.06$), while a separate dimension causally drove the model's choice, with peak intervention effect at layer 19 (95% CI [-4.76, -1.18]; layers 13–21 significant). A model's stated self-assessment can encode cleanly and still fail to explain what it does.

Interpretable multi-agent workflow for metacognitive feedback

Accepted, AIFE 2026 · Under review, IJAIED

- Designed an Evidence–Decision–Feedback (EDF) architecture that turns a learner's written reflection into targeted teaching support. Five stages score the reflection on the Metacognitive Prompting (MP) dimensions, detect the weakest, pedagogically tag why it is weak, select a scaffold, and personalize the feedback to the learner's profile. Each stage records its evidence, so any one can be audited.
- Outlined five design principles for interpretable AI workflows: construct validity, evidence grounding, uncertainty awareness, inference separation, and stakeholder-center intelligibility.
- Traced learner performance through the workflow across successive weeks, from scored reflection to scaffold to their next reflection, where stated intention became and action named at the point of decision (Final Decision, 0.36 to 1.06).
- Discussed what deploying the workflow would require, given that a learner's reflection is untrusted text and can carry hidden instructions. Full **Dual-LLM** quarantine would block that, but it breaks the design, because the learner's exact words have to move forward as evidence; **CaMeL**-style provenance labels solve this by marking those words as data the later stages can read but not obey. The diagnostic record this produces is sensitive, so **attribute-based encryption** would restrict it to the educator, which puts a person at the point of review, override, and moderation.

EDUCATION

Yale University

New Haven, CT

Ph.D. in Neuroscience · Interdepartmental Neuroscience Program

Aug. 2022 – May 2027 (expected)

- Advisor: Philip R. Corlett, Ph.D.
- Selected coursework: *Advances in Large Language Models* (S&DS 617, J. Sekhon) — led the seminar on RLHF, preference modelling, and alignment; *Machine Learning* (BIOS 534).

Emory University

Atlanta, GA

M.S.P.H. in Biostatistics

Aug. 2017 – May 2019

- Advisor: Suprateek Kundu, Ph.D., Thesis: *A Bayesian Latent Scale Brain Network Approach to Conceptualize PTSD*.
- Recipient of the 2019 Biostatistics Thesis Poster Presentation Award.

North Carolina State University

Raleigh, NC

B.S. in Statistics · B.S. in Neurobiology · Magna Cum Laude

Aug. 2012 – May 2016

TECHNICAL SKILLS

Interpretability & Evaluation

Linear probing, logit lens, activation patching and steering, residual-stream analysis, sparse autoencoders (GemmaScope, Neuronpedia)

LLM Systems

Multi-agent workflow design, prompt-injection containment patterns, LLM-as-judge scoring pipelines, PyTorch, Hugging Face Transformers, OpenAI API

Statistical Modelling

Hierarchical Bayesian modelling (Hierarchical Gaussian Filter), reinforcement learning models, mixed-effects and ridge regression, difference-in-differences, latent class analysis, Joint Graphical Lasso

Languages & Tools

Python (NumPy, SciPy, pandas, scikit-learn), R, MATLAB, JavaScript, Bash, Git, LaTeX, AWS, jsPsych, Qualtrics API

PUBLICATIONS

All entries first-authored.

PEER-REVIEWED & ACCEPTED

Inside self-reflection: the geometry in multi-dimensional scoring of metacognitive structure

Suthaharan P, Corlett PR, Ang Y-S. [BlackboxNLP 2026 — ACL Proceedings \(accepted\)](#)

An interpretable AI workflow for metacognitive reflection feedback in digital learning

Suthaharan P, Pang P, Corlett PR, Tan R, Ang Y-S. [AIFE 2026, NIE/NTU Singapore — Extended Summary \(accepted\)](#)

Self-reflection protects behavior from volatile beliefs linked to paranoia

Suthaharan P, Obeso Castiello S, Ang Y-S, Corlett PR. [Computational Psychiatry, 2026](#)

Lesions to the mediodorsal thalamus but not orbitofrontal cortex enhance volatility beliefs linked to paranoia

Suthaharan P, Thompson SL, Rossi-Goldthorpe R, *et al.*, Corlett PR, Chang SWC. [Cell Reports, 2024](#)

Assumed shared belief about conspiracy theories in social networks protects paranoid individuals against distress

Suthaharan P, Corlett PR. [Scientific Reports, 2023](#)

Computational modeling of behavioral tasks: an illustration on a classic reinforcement learning paradigm

Suthaharan P, Corlett PR, Ang Y-S. [The Quantitative Methods for Psychology, 2021](#)

Paranoia and belief updating during the COVID-19 crisis

Suthaharan P, Reed EJ, Leptourgos P, *et al.*, Groman SM, Corlett PR. [Nature Human Behaviour, 2021](#)

UNDER REVIEW

An interpretable artificial intelligence workflow for metacognitive reflection feedback

Suthaharan P, Bhattacharya P, Tan RHL, Corlett PR, Ang Y-S. [International Journal of Artificial Intelligence in Education — submitted 2026](#)

Identifying and tracking paranoia require distinct behavioral signatures

Suthaharan P and Corlett PR. [Nature Mental Health — submitted 2026](#)

INVITED WRITING

Don't Look Behind You

Suthaharan P. [Yale Scientific Magazine, 2024](#)

Pandemic, Paranoia, and the Participant Pool

Suthaharan P. [CloudResearch, 2021](#)

SOFTWARE

- **Closed-loop interpretable metacognitive agentic AI system** — multi-agent EDF pipeline with per-stage auditability. *In development.*
- **LLM-powered metacognitive scorer** — multidimensional scoring over free-text reflection, with probe-based readouts. *GitHub, 2026.*
- **Reinforcement learning task modelling** — simulation and parameter estimation for two-choice RL paradigms. *GitHub, 2021.*

RESEARCH EXPERIENCE

A*STAR, Singapore

Singapore

A*STAR-ARAP Ph.D. Fellow · [Institute of Advanced Intelligence & Computing \(IAIC\)](#)

Feb. 2026 – Present

- Supervised by Yuen-Siang Ang, Ph.D. (SIT), Prasanta Bhattacharya, Ph.D. (A*STAR IAIC) and Philip R. Corlett, Ph.D. (Yale).
- Lead the interpretability workstream: probing, logit lens, and causal activation patching on Gemma-2-2B to test whether behavioural evidence of a model's reasoning corresponds to the internal computation driving it.
- Built the interpretable multi-agent workflow end to end, from multidimensional scoring and diagnostic tagging through just-in-time adaptive scaffolding, under an Evidence–Decision–Feedback framework.
- Authored the design review of safeguards for LLM workflows deployed against untrusted input, covering containment architecture, explanation gating, and human oversight placement.

Yale School of Medicine

New Haven, CT

Research Assistant turned Graduated Researcher · [Department of Psychiatry](#)

Oct. 2019 – Present

- Supervised by Philip R. Corlett, Ph.D.
- Fit Hierarchical Gaussian Filter models to animal and human behavioural data in MATLAB; established cross-species comparability of belief-updating parameters.
- Built and released a large longitudinal behavioural dataset tracking paranoia and belief updating through the COVID-19 pandemic, published in *Nature Human Behaviour*.
- Implemented a difference-in-differences analysis in R to estimate the effect of state policy on paranoia; developed a Qualtrics-based method for eliciting sociocentric network data.

Harvard Medical School

Summer Research Intern · McLean Hospital

Belmont, MA

May 2019 – Aug. 2019

- Supervised by Yuen-Siang Ang, Ph.D. Simulated and estimated two-choice reinforcement learning behaviour in R; implemented gradient-based optimisation to improve parameter recovery.

Emory University

Graduate Researcher · Department of Biostatistics

Atlanta, GA

Aug. 2017 – May 2019

- Supervised by Suprateek Kundu, Ph.D. Built a Bayesian clustering approach using latent class analysis and Dirichlet processes to identify latent subgroups in PTSD symptom data; applied Joint Graphical Lasso to resting-state fMRI to recover connectivity patterns.

Georgia Institute of Technology

Rotation Student · IARPA MICrONS, Department of Biomedical Engineering

Atlanta, GA

May 2018 – Mar. 2019

- Supervised by Eva Dyer, Ph.D. Performed segmentation, alignment, and visualisation of whole-brain imaging data; programmed a multi-modal registration optimiser in MATLAB for 2P calcium and microCT alignment.

TEACHING

Algorithms and Their Societal Implications (CPSC 4640)

Teaching Fellow · Yale University

New Haven, CT

Aug. 2025 – Dec. 2025

- Advised four student groups from problem conception through mathematical formalisation, algorithmic analysis, and final presentation, on projects in algorithmic fairness, differential privacy, and social choice theory.

Developing and Writing Fellowship Proposals (INP 542)

Teaching Fellow · Yale University

New Haven, CT

Jan. 2025 – May 2025

- Assisted Ifat Levy, Ph.D. and Dustin Scheinost, Ph.D. Reviewed Specific Aims and Research Strategy sections for all students and guided the cohort to on-time NIH F31 submission.

INDUSTRY EXPERIENCE

Tetricus Labs Inc.

Consultant

Tuxedo, NY

Jun. 2022 – Aug. 2022

- Designed and implemented the analysis and visualisation layer on the Tetricus platform; curated behavioural datasets for first-pass analysis.

PRA Health Sciences

Data Analyst Intern · Medical Informatics

Raleigh, NC

Aug. 2016 – Aug. 2018

- Provided analytical decision support for clinical trial strategy. Mined proprietary and public data sources (operational systems, claims, EMR, ClinicalTrials.gov, Citeline) and delivered interactive enrolment maps and statistical summaries in R and Tableau to internal and external clients.

SELECTED TALKS & PRESENTATIONS

An interpretable AI workflow for metacognitive reflection feedback in digital learning (*oral*) AIFE · Singapore · Nov. 2026

Inside self-reflection: the geometry in multi-dimensional scoring of metacognitive structure BlackboxNLP · Budapest · Oct. 2026

Low frequency fluctuations in belief-updating identify psychosis risk (*poster*) Society for Neuroscience · Chicago, IL · Oct. 2024

A double dissociation of belief volatility in lesioned monkeys (*invited, oral*) Halassa Lab, Tufts University · Jan. 2024

Belief-updating across species (*oral*) Yale Student Research Talks · New Haven, CT · Jun. 2023

Modeling decision-making using reinforcement learning theory (*oral*) Yale Psychiatry Ribicoff Seminar · Nov. 2021

HONORS & AWARDS

- **2026** ARAP Ph.D. Fellowship, A*STAR Singapore
- **2024** Lykouratzos Distinguished Scholar, Yale University
- **2022** Kavli Institute for Neuroscience Fellow, Yale University
- **2022** COSMOS, Universität Konstanz — selected from 93 global applicants
- **2019** Biostatistics Thesis Poster Presentation Award, Emory University
- **2016** Mu Sigma Rho National Honor Society for Statistics, North Carolina State University
- **2014** NSF Research Experience for Undergraduates, UNC Greensboro