Matt McManus New York, NY — mattmcmanus41@gmail.com — mmcmanus1.github.io/research

Research Summary

I study structure-aware learning for sequential decision-making: building agents that encode known dynamics and causal structure while learning the rest with RL and deep networks. My approach is verification-first—state hypotheses, run ablations with leakage-aware temporal controls, and use mechanistic probes/ interventions to tie internal variables to behavior and uncertainty. I've pursued this theme in three distinct lines of work (not one integrated system): cyber defense, physics-informed inertial navigation, and mechanistic representation analysis. With that lens, I joined Prof. Una-May O'Reilly's CSAIL/ALFA group to frame cyber defense as structured sequential decision-making.

At CSAIL/ALFA (Prof. Una-May O'Reilly), I built a graph-based cyber range to study attacker–defender dynamics as a sequential decision problem. I first added a neuro-symbolic layer on top of RL policies to distill decisions into human-readable rules for audit and mitigation; in head-to-head tests it didn't outperform straightforward RL and added complexity, so we dropped it. We then pivoted to LLM-assisted defense: wiring GPT-3 in as an advisor to the defensive agent for anomaly triage, vulnerability identification, and attack-path planning. In simulation, GPT-3 frequently charted near-optimal paths with checkable rationales, so we kept it as an assistive tool rather than an autonomous controller, validated with hypothesis-driven ablations and leakage-aware temporal controls.

The same principle—encode known structure rather than rely on black-box fitting—motivated my M.Eng. with Prof. Alan Edelman (CSAIL/Julia Lab) in navigation. I developed a physics-informed neural ODE for strapdown inertial navigation in GPS-denied environments, directly embedding kinematic and sensor error structures into the model and training it on simulated and real IMU trajectories. In walk-forward tests, we reduced 3D position RMSE by ∼60% versus a tuned Extended Kalman Filter. I also built a Julia-based simulation/evaluation harness and released CI-backed pipelines. Encoding known physics improved out-of-sample accuracy and made results easier to interpret. [Thesis PDF]

To complement architectural structure, I studied structure in *representations*—what computational models perform internally. In *How Do Transformers "Do" Math?*, we tested whether a transformer trained for linear regression internally represents the intermediate slope w. Linear probes revealed robust encodings of w in hidden states, and reverse probes plus representational interventions that set $w \to w'$ shifted predictions as expected—causal evidence that the model uses a specific internal variable to carry out the computation. [Transformers Math Paper PDF]

Separately, I examined how capacity and regularization shape those internals. In *Low-Complexity Interpolation for Deep Neural Networks*, we biased training toward small-norm, interpolating solutions and observed lower test error and smaller weight norms (consistent with double descent), alongside simpler intermediate features. Encouraging low-complexity solutions made internal computations easier to probe and to regularize. [Low-Complexity DNN Paper PDF]

I carry these principles into industry. At Two Sigma (part-time), I built factor-neutral cross-sectional signals and validated them with leakage-controlled walk-forward tests. At Bridgewater's AIA Labs, I'm focused on LLM reliability—calibration and prompt optimization. Separately, I fixed search indexing in Perplexity's Search API and built a time-aware auditing pipeline to quantify news contamination.

In the future, I want to study how post-training and other policy-tuning methods reshape behavior and uncertainty; design calibration-first objectives and selective prediction so agents state what they know and defer when they don't; develop methods to read and edit the representations that drive decisions using causal tests and interventions; and treat prompt optimization and reflection as objective-driven program search and control rather than ad hoc tweaking. This builds on my prior work in RL framing, mechanistic probes, and evaluation/calibration, and I'll validate across varied settings without being limited to a single domain. The goal is to develop calibrated, steerable systems whose behavior remains understandable, auditable, and improvable as they scale.