Matt McManus

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

Research Summary

I came to research through agents. Building a bot for the MIT Pokerbots Competition as a first-year—and later serving as president—made reinforcement learning concrete and set the tone for my work: frame problems as sequential decisions and integrate *structured priors* so behavior is legible and reliable. That perspective was reinforced by 6.S058 (Representation, Inference, and Reasoning in AI) and 6.S083 (Computational Thinking with Julia), which paired planning under uncertainty with software practices (automatic differentiation, matrix calculus, parallel/GPU) that make results reproducible at scale.

Guided by that agent-centric view, I joined **Prof. Una-May O'Reilly**'s CSAIL/ALFA group and treated cyber defense as sequential decision-making under structure. I built a graph-based cyber-defense simulator, benchmarked RL baselines across network topologies and information regimes with ablations and failure analysis, and explored neuro-symbolic decision layers to distill learned behaviors into human-readable rules. After head-to-head tests, we pivoted when simpler RL matched performance—an informative negative result that clarified when symbolic structure adds value. In parallel, I evaluated LLM-assisted workflows for anomaly triage and attack-path reasoning, where GPT-style models guided a defensive agent's choices.

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 settings, embedding kinematic and sensor-error structure directly in the model and training on simulated and real IMU trajectories. In walk-forward tests, we reduced 3D position RMSE by $\sim$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 computations 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]

Orthogonally, 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 have carried these principles into industry. At Two Sigma as a Quantitative Reasearcher (part-time) I developed factor-neutral cross-sectional signals, designed feature- and learner-level decorrelation (orthogonalization; correlation-penalized loss), and validated with leakage-controlled walk-forward pipelines (rolling normalization; OOS rank IC/IR). At Bridgewater's AIA Labs, I have built *evaluation and calibration systems* for LLM-driven analysis and explore *policy-style prompt optimization.* The work emphasizes leakage-aware temporal controls, calibrated confidence and dispersion/consistency signals, reproducible services (FastAPI/Kubernetes), and measured iteration via offline metrics and lightweight online A/B tests.

Across RL, SciML, and interpretability, the unifying aim is consistent: integrate structured priors with data-driven learning to produce transparent, reliable systems whose internal computations can be probed and steered.