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—watching it adapt against other agents—made reinforcement learning concrete and set the tone for my work: frame problems as sequential decisions with feedback, and inject structure (physics, graphs, symbols, time) so behavior is legible and reliable. 6.S058 (Representation, Inference, and Reasoning in AI) and 6.S083 (Computational Thinking with Julia) reinforced that perspective, pairing planning under uncertainty with software practices—automatic differentiation, matrix calculus, parallel/GPU—that make research reproducible at scale. I later served as president of MIT Pokerbots, which deepened my interest in multi-agent strategy and evaluation.

With **Prof. Una-May O'Reilly** (CSAIL/ALFA), I built a graph-based cyber-defense simulator and cast defense as sequential decision-making. We 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 large-language-model (LLM)–assisted workflows for anomaly triage and attack-path reasoning in the same simulator, where GPT-style models guided a defensive agent's choices. [GPT-3 Cyber Defense PDF]

As an M.Eng. student with **Prof. Alan Edelman** (CSAIL/Julia Lab), I asked how to combine physical priors with learning to improve navigation when GPS is denied. I developed a physics-informed neural ODE for strapdown inertial navigation, embedding kinematics 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 ∼60% versus a tuned Extended Kalman Filter (EKF). I 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]

I also study the mechanisms by which modern models compute. 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; reverse probes and 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.

In separate work on model capacity, "Low-Complexity Interpolation for Deep Neural Networks" biases training toward small-norm, interpolating solutions. We observed lower test error and smaller weight norms (consistent with double descent), alongside simpler intermediate features—evidence that encouraging low-complexity solutions can make internals easier to analyze and regularize.

Industry roles have pushed me to operationalize these ideas. At Two Sigma (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 design evaluation and calibration systems for LLM-driven analysis and frame prompt optimization as RL under explicit temporal/leakage constraints. I have shipped online A/B tests; enforced recency windows and source allowlists; and added dispersion/consistency scoring with Platt-scaled probability outputs. Operationally, these run as FastAPI services on Kubernetes with autoscaling and process-plus-I/O pools, yielding 3-4× higher audit throughput and keeping runs traceable and reproducible.

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