ABSTRACT

Ike Obi, Purdue University, August 2026. Constrained Mediation for Safe Language-Controlled Robot Systems. Major Professors: Byung-Cheol Min and Romila Pradhan.

Recent advances in language-controlled robot systems now allow robots to accept open-ended, free-form natural language task commands. While this capability allows more people to interact with robots, it has also introduced a new safety challenge, wherein language-controlled robot systems now require a safety interface between task specifications and robot task representations. This dissertation establishes constrained mediation as a safety interface for language-controlled robot systems. Constrained mediation in this dissertation, refers to the process of leveraging structured interception, diagnosis, repair, constraint reasoning, and transformation to ensure that task specifications produce only safety-admissible robot task representations. These constrained representations are then synthesized into executable task plans, policies, robot code, and other robot behaviors.

Existing safety methods for language-controlled robot systems typically apply safety constraints during or after generating robot task representations. This dissertation adopts a stricter posture and instead moves safety upstream by applying constrained mediation immediately after the command enters the system and before executable artifacts are synthesized. This posture ensures that downstream planning and control modules receive pre-verified, safety-aware task representations. To realize this objective, this dissertation introduces algorithms and frameworks that that collectively perform task admissibility reasoning, constraint enforcement, consequence anticipation, and command repair prior to synthesis.

To lay the foundation for this research area, the dissertation first introduces SafeGate, a neurosymbolic pre-execution safety architecture that implements a deterministic decision gate and sits at pre-execution and runtime authorization points within LLM-controlled robot systems. SafeGate extracts structured safety-relevant properties from natural language task specifications and applies a deterministic decision gate to authorize, reject, or defer execution. For task specifications that are admitted by the constrained mediation layer, SafeGate further introduces Task Safety Contracts with Z3-based SMT constraint checking for runtime monitoring. Evaluations demonstrates that SafeGate substantially enhances the safety profile of LLM-controlled robot systems compared to existing LLM-based robot safety frameworks.

Recognizing that users need an intuitive pathway to recover from task specification errors, the dissertation further introduces constrained recovery, allowing users to participate in the constrained mediation process. It introduces a graph-based task specification repair framework that transforms defective specifications into safe, executable alternatives that maintain fidelity to the original version. The system uses a neurosymbolic pipeline to first detect safety and feasibility violations, then diagnose causally failure bearing variables over graph-structured task representations, next it generate constrained alternatives, and then validates these candidate repairs symbolically. These generated repair alternatives are then further ranked by balancing task fidelity against intervention cost, favoring repairs that preserve the original specification as much as possible while restoring executability and safety.

Building further on this recovery perspective, this dissertation introduces a constrained preference-learning system for safe robot specification repair. Given a defective task command specification, the system constructs a vetted set of safe repair candidates and learns personalized repair preferences using a Gaussian posterior over interpretable repair features, ensuring that preference learning operates only within a safety-admissible set. By constraining preference learning to a safety-admissible set, this dissertation demonstrates how personalization can be incorporated into the constrained mediation layer without sacrificing safety guarantees.

Empirical studies across all the contributions from this dissertation demonstrated that applying constrained mediation to task specifications before synthesizing executable robot artifacts yields substantial improvements in safety, controllability, recoverability, and task completion success rates of robots in both simulated and real-world environments when compared to existing baselines.