Trustworthy Autonomy Without Human Intervention in Uncertain Domains

Michelle Ho

Department of Aeronautics and Astronautics
Stanford University
496 Lomita Mall, Stanford, CA 94305
mtho@stanford.edu

Abstract

Autonomous systems operating in uncertain environments without human intervention must consider several factors, including safety, reliability, and task success. State-of-the-art methods have made progress in addressing these factors individually, but often fail to unify them for deployment in real-world systems. My dissertation aims to combine methods in planning under uncertainty, failure recovery, and explainability, providing a holistic framework for comprehensive safe autonomy in real-world deployment.

Introduction

Autonomous systems are increasingly deployed in complex, uncertain environments, where they must make their own decisions without human intervention. These decisions have critical implications for safety, reliability, and task success, yet current approaches often address only one isolated aspect of this challenge. For instance, there have been advances in planning under uncertainty, failure detection, and improving planning with large language models. This gap raises the question: how can these capabilities be unified in a framework that enables autonomy to operate reliably across uncertain domains without human oversight?

A primary requirement for trustworthy autonomy is the ability to plan under uncertainty. In the real world, systems cannot assume perfect knowledge of their state, dynamics, or environment, nor can they anticipate every situation they may encounter. Instead, they should consider these uncertainties, along with safety requirements, by planning in the belief space. Robustness to known failures and new anomalies is also essential. In practice, an autonomous system cannot depend on a human operator to intervene every time a failure arises. Instead, the system itself must predict when a failure is imminent and react appropriately to preserve safety and task performance. Transparency is also essential for trustworthy autonomy. Without clear explanations, autonomous planners appear as black boxes, leaving end users unsure why decisions were made, especially when the system is reasoning under multiple sources of uncertainty. This lack of transparency makes it difficult for end users to trust autonomous systems that they did not design themselves.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Trustworthy autonomy requires more than effective performance. Real-world systems must be robust to uncertainties and failures and provide explanations of their behavior. Existing approaches often address one of these challenges in isolation, but few consider them together. The problem, therefore, is how to design autonomy that integrates all three capabilities so that systems can operate reliably across uncertain and complex domains without human oversight.

Approach

My research develops methods for trustworthy autonomy, linking together uncertainty-aware planning, failure anticipation, and explainability. My research so far has explored each of these elements separately to set up their eventual integration into a unified framework.

Belief Space Planning

One common method for planning under uncertainty uses a Partially Observable Markov Decision Process (POMDP), a framework for defining nondeterministic dynamics and partial state observability in a problem (Kochenderfer, Wheeler, and Wray 2022). ρ-POMDPs further introduce belief-dependent rewards (Araya et al. 2010) to explicitly encourage actions that balance between strategically gathering information and completing tasks (Ott, Balaban, and Kochenderfer 2023). Exact planning in POMDPs is generally intractable (Kochenderfer, Wheeler, and Wray 2022), leading to the use of approximate methods. Platt Jr. et al. (2010) proposed belief-iLQR (BiLQR) for planning over full belief-states, which selects actions that reduce future state uncertainty for nonlinear belief dynamics (Platt Jr. et al. 2010). My work extends BiLQR to solve model identification adaptive control (MIAC) (Öreg, Shin, and Tsourdos 2019), or simultaneously controlling the system and learning its dynamics. This extension supports trustworthy autonomy by allowing planning to adapt online to imperfect models. My ongoing work extends this idea to incorporate chance constraints, so even in uncertain conditions, safety requirements are satisfied with high probability. The goal is to successfully plan with higher fidelity models that more closely resemble real-world conditions.

Failure Anticipation

Detecting and responding to failures has long been necessary for systems in uncertain environments. The two main failure detection classes are failure classification, which identifies failures that the system has previously seen and knows how to correct or recover from (Costa et al. 2019), and anomaly detection, which identifies deviations from nominal data without explicitly labeled failure data (Chandola, Banerjee, and Kumar 2009). Trustworthy autonomy requires both. While we cannot assume we can train the system on every possible failure, it must also know how to respond appropriately when a common failure case is anticipated. Recent work has applied conformal prediction techniques to anomaly detection, allowing for failure anticipation just from success data (Xu et al. 2025). Building on this, my work extended this anomaly detection method to multi-camera-view settings by leveraging world-model prediction error. In a follow-up work, I added failure classification alongside the anomaly detection, creating a unified approach that identifies both familiar failures and novel ones, on a mobile quadruped. The next step is to integrate this framework with planning, enabling systems to recover appropriately after anticipating a specific failure.

Large Language Model Explainability

Interpretability is a barrier for trusting AI in safety-critical domains (Doshi-Velez and Kim 2017). Trajectories designed with AI often require extensive human verification before deployment. This task could be automated with large language models (LLMs) by leveraging their natural text generation and inference capabilities. Previous approaches that combine LLMs and planning primarily focus on improving performance rather than explaining plans (Ding et al. 2023). My ongoing work addresses this gap by translating belief trees from Monte Carlo Tree Search (MCTS), a common method for planning under uncertainty (Kochenderfer, Wheeler, and Wray 2022), into structured LLM text inputs. A user can ask why a specific action was chosen in a trajectory, what trade-offs it involved, or why other paths are suboptimal. The agent is demonstrated on a rover path planning problem, where uncertainty, risk, and mission constraints make transparency essential. Next steps involve improving accuracy through LLM ensembles and enabling the agent to assess failures and propose recovery strategies.

Research Plan

My thesis aims to advance trustworthy autonomy by integrating planning under uncertainty, failure anticipation, and natural-language explanations, to enable reliable operation in uncertain environments. I have completed three works towards this goal. First, I extended BiLQR to MIAC with a ρ-POMDP formulation, balancing reducing uncertainty and improving performance. Second, I developed a world model anomaly detection framework, showing that world models can detect failures without explicit failure labels, for multiple synced camera views. Third, I extended this idea to a unified anomaly detection and failure classification framework on a mobile, legged robot, laying the groundwork for executing specific intervention strategies.

By the consortium, I will evaluate methods for integrating chance constraints into belief-space planning methods, such as BiLQR. Then, for my LLM explainability work, I will conduct ensemble testing across multiple LLMs to reduce hallucinations via agreement, extending beyond initial GPT-based feasibility results. Lastly, I am exploring how vision-language-action models can suggest semantic failure recovery, rather than using pixel-level pattern recognition.

My research will then shift toward linking those threads. By summer 2026, I plan to pursue one of the following: (1) leveraging semantics from LLMs for failure anticipation, where an LLM could recommend appropriate intervention strategies once a class of failure is identified; or (2) incorporating world model failure anticipation and recovery into belief-space planning, to use conformal prediction scores to guide corrective actions that reduce deviation from successful behavior. By winter 2027, I expect to explore the other idea and finally expand the explainability framework to the entire pipeline by fall 2027. This timeline will lead to dissertation writing between fall 2027 and spring 2028.

References

Araya, M.; Buffet, O.; Thomas, V.; and Charpillet, F. 2010. A POMDP Extension with belief-dependent rewards. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3): 1–58.

Costa, M. A.; Wullt, B.; Norrlöf, M.; and Gunnarsson, S. 2019. Failure detection in robotic arms using statistical modeling, machine learning and hybrid gradient boosting. *Measurement*, 146: 425–436.

Ding, Y.; Zhang, X.; Paxton, C.; and Zhang, S. 2023. Task and motion planning with large language models for object rearrangement. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2086–2092.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv*.

Kochenderfer, M. J.; Wheeler, T. A.; and Wray, K. H. 2022. *Algorithms for Decision Making*. MIT Press.

Ott, J.; Balaban, E.; and Kochenderfer, M. J. 2023. Sequential Bayesian optimization for adaptive informative path planning with multimodal sensing. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Platt Jr., R.; Tedrake, R.; Kaelbling, L.; and Lozano-Perez, T. 2010. Belief space planning assuming maximum likelihood observations. In *Robotics: Science and Systems*.

Xu, C.; Nguyen, T. K.; Dixon, E.; Rodriguez, C.; Lee, R.; Miller, P.; Shah, P.; Ambrus, R.; Nishimura, H.; and Itkina, M. 2025. Can we detect failures without failure data?: Uncertainty-aware runtime failure detection for imitation learning policies. In *Robotics: Science and Systems*.

Öreg, Z.; Shin, H.-S.; and Tsourdos, A. 2019. Model identification adaptive control: implementation case studies for a high manoeuvrability aircraft. In *Mediterranean Conference on Control and Automation*, 559–564.