Manipulating Neural Path Planners via Slight Perturbations

Zikang Xiong and Suresh Jagannathan

***Abstract*—Data-driven neural path planners are attracting increasing interest in the robotics community. However, their neural network components typically come as black boxes, obscuring their underlying decision-making processes. Their black-box nature exposes them to the risk of being compromised via the insertion of hidden malicious behaviors. For example, an attacker may hide behaviors that, when triggered, hijack a delivery robot by guiding it to a specific (albeit wrong) destination, trapping it in a predefined region, or inducing unnecessary energy expenditure by causing the robot to repeatedly circle a region. In this letter, we propose a novel approach to specify and inject a range of hidden malicious behaviors, known as backdoors, into neural path planners. Our approach provides a concise but flexible way to define these behaviors, and we show that hidden behaviors can be triggered by slight perturbations (e.g., inserting a tiny unnoticeable object), that can nonetheless significantly compromise their integrity. We also discuss potential techniques to identify these backdoors aimed at alleviating such risks. We demonstrate our approach on both sampling-based and search-based neural path planners.**

***Index Terms*—Deep learning methods, integrated planning and learning, motion and path planning, robot safety.**

I. Introduction

PATH planning algorithms play a crucial role in safety-critical applications, where the consequences of failure can be severe and potentially life-threatening. These applications include autonomous vehicles, where the quality of path plans directly correlates to vehicle safety [1], [2], robotic arm manip-

example, a classifier could wrongly identify a stop sign as a green light when an undetectable trigger is added to an image. Despite the extensive study of backdoor attacks in computer vision [15], [16] and natural language processing [17] to induce misclassifications, they present distinct challenges in path planning problems. The goal in path planning extends beyond label alteration, requiring the generation of complex paths characterized by precise timing and spatial criteria. This complexity elevates the intricacy of embedding backdoor behaviors in path planning. Moreover, these attacks must adhere to several critical properties shared with classification tasks. First, the attacks must be easy to trigger with only slight changes to the environment. Second, they need to be persistent even when the input varies. Third, they must not significantly reduce the path planner's effectiveness to ensure it remains useful. Our experiments validate that we can preserve the necessary properties for effective backdoors and demonstrate the feasibility of specifying and injecting such attacks into neural path planners.

Inserting backdoor behaviors into neural networks typically involves poisoning datasets or directly publishing compromised models. These two types of attacks are a rising source of concern. For example, many robotics datasets are now open to the public with anyone able to contribute to them [18]. Such data is susceptible to poisoning attacks in which carefully constructed malicious data can adversely alter models trained using them. Similarly, it is increasingly common for models to be published