Table 3: JSC vs. Q-learning with Error Injection (.05 error rate)

Training steps	JSC			Normal		
	Crash	Fall behind	Steady	Crash	Fall behind	Steady
1,000	3	12539	306	10950	1745	153
10,000	7	12502	339	10546	2215	87
100,000	5	12359	484	10561	2242	45

with a slightly modified algorithm. The modified algorithm uses the distance between the current state and the modeled environment as an objective function during speculation; when the system leaves the modeled state space, the reinforcement learning agent optimizes for returning to the modeled portion of the state space. We call this approach JSCQ.

Table 4: Crashing states for JSC and JSCQ control

Perturbation	JSC	JSCQ
5%	8	2
25%	18	16
50%	41	34

For small positional perturbations, JSCQ and JSC perform equally well. However, we found that as the positional perturbation increases, the modified algorithm begins to outperform the original algorithm. Table 4 presents these results.

Related Work

Safe Artificial Intelligence is an emerging area of interest, but there is a rich history of research on safe control in the absence of perfect models.

There are a myriad of approaches toward safe reinforcement learning that do not take advantage of formal verification. Many of these approaches are summarized in (García and Fernández 2015), who decompose these approaches into two broad categories: modification of the optimality criterion, and modification of the exploration process. In this section, we compare our approach to these approaches, following García and Fernández's taxonomy.

The primary novel contribution of our work, compared to this body of literature, is two-fold. First, we leverage hybrid systems verification results and runtime monitor synthesis to appropriately sandbox the exploration process, instead of relying on more ad-hoc sources of knowledge about how to act safely. The chain of evidence transfers from a high-level model to runtime monitors and ultimately to the reinforcement learning process via the theorems presented in this paper. Second, we distinguish between *optimizing among known safe policy options* and *speculating about portions of the state space that are not a priori modeled*. This distinction is crucial to determine what level of speculation should be allowed, and when.

When compared to existing approaches to reinforcement learning, our approach either 1) suggests a way to strengthen the existing approach by incorporating not just a known-safe policy but a *formally verified* safe policy; or 2) is possibly compositional with the existing approach (by further modifying our exploration process to perform more robust decision making when the model monitor is already violated).

Many approaches to safe reinforcement learning work by constraining the optimality criterion used in reinforcement learning. The idea is simple – instead of selecting from the entire policy space, the agent may only choose from a set of control actions that are a priori known to be safe (García and Fernández 2015). There are several approaches with this basic flavor (Kadota, Kurano, and Yasuda 2006; Geibel 2006; Moldovan and Abbeel 2012). Our primary contribution, relative to this work, is that we retain formal verification results for the restricted policy space through the use of provably correct runtime monitors, and allow speculation when policy space restriction becomes unjustified due to deviation from modeling assumptions. Other approaches do not provide such strong results and do not relax constraints when modeling assumptions are violated, but do perform more sophisticated approaches toward learning (we simply perform naive Q-learning on a discretized state space). Incorporating these more sophisticated approaches in a way that retains the safety theorems presented in this paper is left as promising future work.

Another approach toward safe reinforcement learning adopts worst case criterion (Heger 1994) or risk-sensitive criterion (Tamar, Xu, and Mannor 2013; Nilim and Ghaoui 2005), in which the optimization criterion is modified to reflect safety concerns.

Another set of approaches initialize the learner with some initial knowledge, with the goal of directing policy exploration away from unsafe states (Driessens and Dzeroski 2004). Our approach is analogous – for example, in the context of teacher/learner reinforcement learning, our nondeterministic controls could be thought of as an infinite set of demonstrations. The problem in both cases is how to safely control in cases where no demonstration is provided. The primary strength of our approach is that we leverage formal verification, and preserve these results during exploration. We also believe it is important to distinguish between optimization within a known safe policy space, and exploration of inherently speculative control options.

Another approach toward safe reinforcement learning involves analysis of the policy constructed from a learning process (Katz et al. 2017). These approaches are appropriate when the learning phase is not safety-critical, but not appropriate when the system must behave safely *while learning*. We do both and also work with verified models, instead of depending upon conjectures or assumptions about the model. Our approach is also computationally attractive when compared to these approaches, because we enforce