From Cold Start to Critical: Formal Synthesis of Hybrid Controllers

PI: Dane A. Sabo
dane.sabo@pitt.edu

Advisor: Dr. Daniel G. Cole
dgcole@pitt.edu

Wednesday 15th October, 2025

1 Goals and Outcomes

The goal of this research is to develop a methodology for creating autonomous hybrid control systems with mathematical guarantees of safe and correct behavior.

Nuclear power plants require the highest levels of control system reliability, where failures can result in significant economic losses, service interruptions, or radiological release. Currently, nuclear plant operations rely on extensively trained human operators who follow detailed written procedures and strict regulatory requirements to manage reactor control. These operators make critical decisions about when to switch between different control modes based on their interpretation of plant conditions and procedural guidance. However, this reliance on human operators prevents the introduction of autonomous control capabilities and creates a fundamental economic challenge for next-generation reactor designs. Emerging technologies like small modular reactors face significantly higher per-megawatt staffing costs than conventional plants, threatening their economic viability. What is needed is a way to create autonomous control systems that can safely manage complex operational sequences with the same level of assurance as human-operated systems, but without requiring constant human supervision.

To address this need, we will combine formal methods from computer science with control theory to build hybrid control systems that are correct by construction. Hybrid systems use discrete logic to switch between continuous control modes, similar to how operators change control strategies. Existing formal methods can generate provably correct switching logic from written requirements, but they cannot handle the continuous dynamics that occur during transitions between modes. Meanwhile, traditional control theory can verify continuous behavior but lacks tools for proving correctness of discrete switching decisions. By synthesizing discrete mode transitions directly from written operating procedures and verifying continuous behavior between transitions, we can create hybrid control systems with end-to-end correctness guarantees. If we can formalize existing procedures into logical specifications and verify that continuous dynamics satisfy transition requirements, then we can build autonomous controllers that are provably free from design defects. This approach will enable autonomous control in nuclear power plants while maintaining the high safety standards required by the industry. This work is conducted within the University of Pittsburgh Cyber Energy Center, which provides access to industry collaboration and Emerson control hardware, ensuring that solutions developed are aligned with practical implementation requirements.

If this research is successful, we will be able to do the following:

1. **Translate written procedures into verified control logic.** We will develop a methodology for converting existing written operating procedures into formal specifications that can be automatically synthesized into discrete control logic. This process will use structured intermediate representations to bridge natural language procedures and mathematical logic. Control system engineers will be able to generate verified mode-switching controllers directly from regulatory procedures without requiring expertise in formal methods, reducing the barrier to creating high-assurance control systems.
2. **Verify continuous control behavior across mode transitions.** We will establish methods for analyzing continuous control modes to ensure they satisfy the discrete transition requirements. Using a combination of classical control theory for linear systems and reachability analysis for nonlinear dynamics, we will verify that each continuous mode can safely reach its intended transitions. Engineers will be able to design continuous controllers using

standard practices while iterating to ensure broader system correctness, proving that mode transitions occur safely and at the right times.

3. **Demonstrate autonomous reactor startup control with safety guarantees.** We will apply this methodology to develop an autonomous controller for nuclear reactor startup procedures, implementing it on a small modular reactor simulation using industry-standard control hardware. This demonstration will prove correctness across multiple coordinated control modes from cold shutdown through criticality to power operation. We will provide evidence that autonomous hybrid control can be realized in the nuclear industry with current control equipment, establishing a path toward reducing operator staffing requirements while maintaining safety.

The innovation in this work is the unification of discrete synthesis and continuous verification to enable end-to-end correctness guarantees for hybrid systems. If successful, control engineers will be able to create autonomous controllers from existing procedures with mathematical proof of correct behavior. High-assurance autonomous control will become practical for safety-critical applications. This capability is essential for the economic viability of next-generation nuclear power. Small modular reactors represent a promising solution to growing energy demands, but their success depends on reducing per-megawatt operating costs through increased autonomy. This research will provide the tools to achieve that autonomy while maintaining the exceptional safety record required by the nuclear industry.

2 State of the Art and Limits of Current Practice

Nuclear reactor control represents a quintessential hybrid cyber-physical system. Continuous physical plant dynamics—neutron kinetics, thermal-hydraulics, heat transfer—interact with discrete control logic—mode transitions, trip decisions, valve states. Yet **formal hybrid control synthesis methods remain largely unapplied** to this safety-critical domain. This gap persists despite compelling evidence: human error contributes to **70–80% of all nuclear incidents** [1–3] even after four decades of improvements in training, procedures, and automation.

Current reactor control practices lack the mathematical guarantees that formal verification could provide. Recent efforts to apply formal methods—such as the HARDENS project—have addressed only discrete control logic without considering continuous reactor dynamics or experimental validation. This section examines three critical areas: existing reactor control practices and their fundamental limitations, the persistent impact of human factors in nuclear safety incidents, and pioneering formal methods efforts that demonstrate both the promise and current limitations of rigorous digital engineering for nuclear systems. Together, these areas reveal a clear research imperative: to develop mathematically verified hybrid controllers that provide safety guarantees across both continuous plant dynamics and discrete control logic while addressing the reliability limitations inherent in human-in-the-loop control.

2.1 Current Reactor Control Practices

Nuclear reactor control in the United States and globally relies on a carefully orchestrated combination of human operators, written procedures, automated safety systems, and increasingly digital instrumentation and control (I&C) systems. Understanding current practices—and their limitations—provides essential context for motivating formal hybrid control synthesis.

2.1.1 Human Operators Retain Ultimate Decision Authority

Current generation nuclear power plants employ **3,600+ active NRC-licensed reactor operators** in the United States, divided into Reactor Operators (ROs) who manipulate reactor controls and Senior Reactor Operators (SROs) who direct plant operations and serve as shift supervisors [4]. These operators work in control rooms featuring mixed analog and digital displays, enhanced by Safety Parameter Display Systems (SPDS) mandated after the Three Mile Island accident. Staffing typically requires **2–4 operators per shift** for current generation plants, though advanced designs like NuScale have demonstrated that operations can be conducted with as few as three operators.

The role of human operators is paradoxically both critical and problematic. Operators hold legal authority under 10 CFR Part 55 to make critical decisions including departing from normal regulations during emergencies—a necessity for handling unforeseen scenarios but also a source of risk. The Three Mile Island accident demonstrated how "combination of personnel error, design deficiencies, and component failures" led to partial meltdown when operators "misread confusing and contradictory readings and shut off the emergency water system" [5]. The President's Commission on TMI identified a fundamental ambiguity: placing "responsibility and accountability for safe power plant operations...on the licensee in all circumstances" without formal verification that operators can fulfill this responsibility under all conditions [5]. This tension between operational flexibility and safety assurance remains unresolved in current practice.

Advanced designs attempt to reduce operator burden through passive safety features and increased automation. NuScale's Small Modular Reactor design requires **no operator actions for 72 hours** following design-basis accidents and only two operator actions for beyond-design-basis events. However, even these advanced designs retain human operators for strategic decisions, procedure implementation, and override authority—preserving the human reliability challenges documented over four decades.

2.1.2 Operating Procedures Lack Formal Verification

Nuclear plant procedures exist in a hierarchy: normal operating procedures for routine evolutions, abnormal operating procedures for off-normal conditions, Emergency Operating Procedures (EOPs) for design-basis accidents, Severe Accident Management Guidelines (SAMGs) for beyond-design-basis events, and Extensive Damage Mitigation Guidelines (EDMGs) for catastrophic damage scenarios. These procedures must comply with 10 CFR 50.34(b)(6)(ii) and are developed using guidance from NUREG-0899 [6], but their development process relies fundamentally on expert judgment and simulator validation rather than formal verification.

EOPs adopted a symptom-based approach following TMI, allowing operators to respond to plant conditions without first diagnosing root causes—a significant improvement over earlier event-based procedures. The BWR Owners' Group completed Revision 3 of integrated Emergency Procedure Guidelines/Severe Accident Guidelines in 2013, representing the current state of the art in procedure development. Procedures undergo technical evaluation, simulator validation testing, and biennial review as part of operator requalification under 10 CFR 55.59 [4].

Despite these rigorous development processes, **procedures fundamentally lack formal verification of key safety properties**. There is no mathematical proof that procedures cover all possible plant states, that required actions can be completed within available timeframes under all scenarios, or that transitions between procedure sets maintain safety invariants. As the IAEA notes in TECDOC-1580 [7], "Most subsequent investigations identify internal and external industry oper-

ating experience that, if applied effectively, would have prevented the event"—a pattern suggesting that current procedure development methods cannot guarantee completeness.

LIMITATION: *Procedures lack formal verification of correctness and completeness.* Current procedure development relies on expert judgment and simulator validation. No mathematical proof exists that procedures cover all possible plant states, that required actions can be completed within available timeframes, or that transitions between procedure sets maintain safety invariants. Paper-based procedures cannot adapt to novel combinations of failures, and even computer-based procedure systems lack the formal guarantees that automated reasoning could provide.

2.1.3 Control Mode Transitions Lack Formal Safety Verification

Nuclear plants operate with multiple control modes: automatic control where the reactor control system maintains target parameters through continuous rod adjustment, manual control where operators directly manipulate control rods, and various intermediate modes. In typical PWR operation, the reactor control system automatically maintains floating average temperature, compensating for xenon effects and fuel burnup at rates limited to approximately 5% power per minute. Safety systems operate with high automation—Reactor Protection Systems trip automatically on safety signals with millisecond response times, and Engineered Safety Features actuate automatically on accident signals without operator action required.

The decision to transition between control modes relies on operator judgment informed by plant stability, equipment availability, procedural requirements, and safety margins. However, current practice lacks formal verification that mode transitions maintain safety properties across all possible plant states. As Žerovnik et al. observe [8], "Manual control may be demanded in nuclear power plants due to safety protocols. However, it may not be convenient in load-following regimes with frequent load changes"—highlighting the tension between operational flexibility and formal safety assurance.

Research by Jo et al. [9] reveals a concerning trade-off: "using procedures at high level of automation enables favorable operational performance with decreased mental workload; however, operator's situation awareness is decreased." This automation paradox—where increasing automation reduces errors from workload but increases errors from reduced vigilance—has been empirically demonstrated but not formally optimized. Operators may experience mode confusion, losing track of which control mode is active during complex scenarios.

LIMITATION: *Mode transitions lack formal safety verification.* No formal proof exists that all mode transitions preserve safety invariants across the hybrid state space of continuous plant dynamics and discrete control logic. The automation paradox trade-off between reduced workload and reduced situation awareness has never been formally optimized with mathematical guarantees about the resulting reliability.

2.1.4 Current Automation Reveals the Hybrid Dynamics Challenge

Approximately **40% of the world's operating reactors** [10] have undergone some digital I&C upgrades, with 90% of digital implementations representing modernization of existing analog systems. All reactors beginning construction after 1990 incorporate digital I&C components, with Asia leading adoption.

The current division between automated and human-controlled functions reveals the fundamental challenge of hybrid control. **Highly automated systems** handle reactor protection (automatic trip on safety parameters), emergency core cooling actuation, containment isolation, and

basic process control. **Human operators retain control** of strategic decision-making (power level changes, startup/shutdown sequences, mode transitions), procedure implementation (emergency response strategy selection), override authority, and assessment and diagnosis of beyond-design-basis events.

Emerging technologies include deep reinforcement learning for autonomous control and Long Short-Term Memory networks for safety system control. Lee et al. demonstrated [11] that autonomous LSTM-based control achieved **performance superior to automation-plus-human-control** in simulated loss-of-coolant and steam generator tube rupture scenarios. Yet even these advanced autonomous control approaches lack formal verification, and as IEEE research documented [12], "Introducing I&C hardware failure modes to formal models comes at significant computational cost...state space explosion and prohibitively long processing times."

LIMITATION: *Current practice treats continuous plant dynamics and discrete control logic separately.* No application of hybrid control theory exists that could provide mathematical guarantees across mode transitions, verify timing properties formally, or optimize the automation-human interaction trade-off with provable safety bounds.

2.2 Human Factors in Nuclear Accidents

The persistent role of human error in nuclear safety incidents, despite decades of improvements in training and procedures, provides perhaps the most compelling motivation for formal automated control with mathematical safety guarantees.

2.2.1 Human Error Dominates Nuclear Incident Causation

Multiple independent analyses converge on a striking statistic: **70–80% of all nuclear power plant events are attributed to human error** versus approximately 20% to equipment failures [1, 2]. More significantly, the International Atomic Energy Agency concluded that "human error was the root cause of all severe accidents at nuclear power plants"—a categorical statement spanning Three Mile Island, Chernobyl, and Fukushima Daiichi [13].

A detailed analysis of 190 events at Chinese nuclear power plants from 2007–2020 by Wang et al. [3] found that 53% involved active errors while 92% were associated with latent errors—organizational and systemic weaknesses that create conditions for failure. Lloyd Dumas's study [14] found approximately 80% of incidents at 10 nuclear centers stemmed from worker error or poor procedures, with roughly 70% from latent organizational weaknesses and 30% from individual worker actions.

The persistence of this 70–80% human error contribution despite **four decades of continuous improvements** in operator training, control room design, procedures, and human factors engineering suggests fundamental cognitive limitations rather than remediable deficiencies.

2.2.2 Three Mile Island Revealed Critical Human-Automation Interaction Failures

The Three Mile Island Unit 2 accident on March 28, 1979 remains the definitive case study in human factors failures in nuclear operations. The accident began at 4:00 AM with a routine feedwater pump trip, escalating when a pressure-operated relief valve (PORV) stuck open—draining reactor coolant—but control room instrumentation showed only whether the valve had been commanded to close, not whether it actually closed. When Emergency Core Cooling System pumps automatically activated as designed, **operators made the fateful decision to shut them down** based on their incorrect assessment of plant conditions.

President's Commission chairman John Kemeny documented [5] how operators faced more than 100 simultaneous alarms, overwhelming their cognitive capacity. The core suffered partial meltdown with **44% of the fuel melting** before the situation was stabilized.

Quantitative risk analysis revealed the magnitude of failure in existing safety assessment methods: the actual core damage probability was approximately **5% per year** while Probabilistic Risk Assessment had predicted 0.01% per year—a **500-fold underestimation**. This dramatic failure demonstrated that human reliability could not be adequately assessed through expert judgment and historical data alone.

2.2.3 Human Reliability Analysis Documents Fundamental Cognitive Limitations

Human Reliability Analysis (HRA) methods developed over four decades quantify human error probabilities and performance shaping factors. The SPAR-H method [15] represents current best practice, providing nominal Human Error Probabilities (HEPs) of **0.01 (1%) for diagnosis tasks** and **0.001 (0.1%) for action tasks** under optimal conditions.

However, these nominal error rates degrade dramatically under realistic accident conditions: inadequate available time increases HEP by **10-fold**, extreme stress by **5-fold**, high complexity by **5-fold**, missing procedures by **50-fold**, and poor ergonomics by **50-fold**. Under combined adverse conditions typical of severe accidents, human error probabilities can approach **0.1 to 1.0 (10% to 100%)**—essentially guaranteed failure for complex diagnosis tasks [16].

Rasmussen's influential 1983 taxonomy [17] divides human errors into skill-based (highly practiced responses, HEP 10^{-3} to 10^{-4}), rule-based (following procedures, HEP 10^{-2} to 10^{-1}), and knowledge-based (novel problem solving, HEP 10^{-1} to 1). Severe accidents inherently require knowledge-based responses where human reliability is lowest. Miller's classic 1956 finding [18] that working memory capacity is limited to **7±2 chunks** explains why Three Mile Island's 100+ simultaneous alarms exceeded operators' processing capacity.

LIMITATION: *Human factors impose fundamental reliability limits that cannot be overcome through training alone.* Response time limitations constrain human effectiveness—reactor protection systems must respond in milliseconds, **100–1000 times faster than human operators**. Cognitive biases systematically distort judgment: confirmation bias, overconfidence, and anchoring bias are inherent features of human cognition, not individual failings [19]. The persistent 70–80% human error contribution despite four decades of improvements demonstrates that these limitations are **fundamental rather than remediable**.

2.3 HARDENS: Discrete Control with Gaps in Hybrid Dynamics

The High Assurance Rigorous Digital Engineering for Nuclear Safety (HARDENS) project, completed by Galois, Inc. for the U.S. Nuclear Regulatory Commission in 2022, represents the most advanced application of formal methods to nuclear reactor control systems to date—and simultaneously reveals the critical gaps that remain.

2.3.1 Rigorous Digital Engineering Demonstrated Feasibility

HARDENS aimed to address the nuclear industry's fundamental dilemma: existing U.S. nuclear control rooms rely on analog technologies from the 1950s–60s, making construction costs exceed $500 million and timelines stretch to decades. The NRC contracted Galois to demonstrate that Model-Based Systems Engineering and formal methods could design, verify, and implement a complex protection system meeting regulatory criteria at a fraction of typical cost.

The project delivered far beyond its scope, creating what Galois describes as "the world's most advanced, high-assurance protection system demonstrator." Completed in **nine months at a tiny fraction of typical control system costs** [20], the project produced a complete Reactor Trip System (RTS) implementation with full traceability from NRC Request for Proposals and IEEE standards through formal architecture specifications to formally verified binaries and hardware running on FPGA demonstrator boards.

Principal Investigator Joseph Kiniry led the team in applying Galois's Rigorous Digital Engineering methodology combining model-based engineering, digital twins with measurable fidelity, and applied formal methods. The approach integrates multiple abstraction levels—from semi-formal natural language requirements through formal specifications to verified implementations—all maintained as integrated artifacts rather than separate documentation prone to divergence.

2.3.2 Comprehensive Formal Methods Toolkit Provided Verification

HARDENS employed an impressive array of formal methods tools and techniques across the verification hierarchy. High-level specifications used Lando, SysMLv2, and FRET (NASA JPL's Formal Requirements Elicitation Tool) to capture stakeholder requirements, domain engineering, certification requirements, and safety requirements. Requirements were formally analyzed for **consistency, completeness, and realizability** using SAT and SMT solvers—verification that current procedure development methods lack.

Executable formal models employed Cryptol to create an executable behavioral model of the entire RTS including all subsystems, components, and formal digital twin models of sensors, actuators, and compute infrastructure. Automatic code synthesis generated formally verifiable C implementations and System Verilog hardware implementations directly from Cryptol models—eliminating the traditional gap between specification and implementation where errors commonly arise.

Formal verification tools included SAW (Software Analysis Workbench) for proving equivalence between models and implementations, Frama-C for C code verification, and Yosys for hardware verification. HARDENS verified both automatically synthesized and hand-written implementations against their models and against each other, providing redundant assurance paths.

This multi-layered verification approach represents a quantum leap beyond current nuclear I&C verification practices, which rely primarily on testing and simulation. HARDENS demonstrated that **complete formal verification from requirements to implementation is technically feasible** for safety-critical nuclear control systems.

2.3.3 Critical Limitation: Discrete Control Logic Only

Despite its impressive accomplishments, HARDENS has a fundamental limitation directly relevant to hybrid control synthesis: **the project addressed only discrete digital control logic without modeling or verifying continuous reactor dynamics**. The Reactor Trip System specification and formal verification covered discrete state transitions (trip/no-trip decisions), digital sensor input processing through discrete logic, and discrete actuation outputs (reactor trip commands). The system correctly implements the digital control logic for reactor protection with mathematical guarantees.

However, the project did not address continuous dynamics of nuclear reactor physics including neutron kinetics, thermal-hydraulics, xenon oscillations, fuel temperature feedback, coolant flow dynamics, and heat transfer—all governed by continuous differential equations. Real reac-

tor safety depends on the interaction between continuous processes (temperature, pressure, neutron flux evolving according to differential equations) and discrete control decisions (trip/no-trip, valve open/close, pump on/off). HARDENS verified the discrete controller in isolation but not the closed-loop hybrid system behavior.

LIMITATION: *HARDENS addressed discrete control logic without continuous dynamics or hybrid system verification.* Hybrid automata, differential dynamic logic, or similar hybrid systems formalisms would be required to specify and verify properties like "the controller maintains core temperature below safety limits under all possible disturbances"—a property that inherently spans continuous and discrete dynamics. Verifying discrete control logic alone provides no guarantee that the closed-loop system exhibits desired continuous behavior such as stability, convergence to setpoints, or maintained safety margins.

2.3.4 Experimental Validation Gap Limits Technology Readiness

The second critical limitation is **absence of experimental validation** in actual nuclear facilities or realistic operational environments. HARDENS produced a demonstrator system at Technology Readiness Level 3–4 (analytical proof of concept with laboratory breadboard validation) rather than a deployment-ready system validated through extended operational testing. The NRC Final Report explicitly notes [20]: "All material is considered in development and not a finalized product" and "The demonstration of its technical soundness was to be at a level consistent with satisfaction of the current regulatory criteria, although with no explicit demonstration of how regulatory requirements are met."

The project did not include deployment in actual nuclear facilities, testing with real reactor systems under operational conditions, side-by-side validation with operational analog RTS systems, systematic failure mode testing (radiation effects, electromagnetic interference, temperature extremes), actual NRC licensing review, or human factors validation with licensed nuclear operators in realistic control room scenarios.

LIMITATION: *HARDENS achieved TRL 3–4 without experimental validation.* While formal verification provides mathematical correctness guarantees for the implemented discrete logic, the gap between formal verification and actual system deployment involves myriad practical considerations: integration with legacy systems, long-term reliability under harsh environments, human-system interaction in realistic operational contexts, and regulatory acceptance of formal methods as primary assurance evidence.

2.4 Research Imperative: Formal Hybrid Control Synthesis

Three converging lines of evidence establish an urgent research imperative for formal hybrid control synthesis applied to nuclear reactor systems.

Current reactor control practices reveal fundamental gaps in verification. Procedures lack mathematical proofs of completeness or timing adequacy. Mode transitions preserve safety properties only informally. Operator decision-making relies on training rather than verified algorithms. The divide between continuous plant dynamics and discrete control logic has never been bridged with formal methods. Despite extensive regulatory frameworks developed over six decades, **no mathematical guarantees exist** that current control approaches maintain safety under all possible scenarios.

Human factors in nuclear accidents demonstrate that human error contributes to 70–80% of nuclear incidents despite four decades of systematic improvements. The IAEA's categorical state-

ment that "human error was the root cause of all severe accidents" reveals fundamental cognitive limitations: working memory capacity of 7 ± 2 chunks, response times of seconds to minutes versus milliseconds required, cognitive biases immune to training, stress-induced performance degradation. Human Reliability Analysis methods document error probabilities of 0.001–0.01 under optimal conditions degrading to 0.1–1.0 under realistic accident conditions. These limitations **cannot be overcome through human factors improvements alone**.

The HARDENS project proved that formal verification is technically feasible and economically viable for nuclear control systems, achieving complete verification from requirements to implementation in nine months at a fraction of typical costs. However, HARDENS addressed only discrete control logic without considering continuous reactor dynamics or hybrid system verification, and the demonstrator achieved only TRL 3–4 without experimental validation in realistic nuclear environments. These limitations directly define the research frontier: **formal synthesis of hybrid controllers that provide mathematical safety guarantees across both continuous plant dynamics and discrete control logic**.

The research opportunity is clear. Nuclear reactors are quintessential hybrid cyber-physical systems where continuous neutron kinetics, thermal-hydraulics, and heat transfer interact with discrete control mode decisions, trip logic, and valve states. Current practice treats these domains separately—reactor physics analyzed with simulation, control logic verified through testing, human operators expected to integrate everything through procedures. **Hybrid control synthesis offers the possibility of unified formal treatment** where controllers are automatically generated from high-level safety specifications with mathematical proofs that guarantee safe operation across all modes, all plant states, and all credible disturbances.

Recent advances in hybrid systems theory—including reachability analysis, barrier certificates, counterexample-guided inductive synthesis, and satisfiability modulo theories for hybrid systems—provide the theoretical foundation. Computational advances enable verification of systems with continuous state spaces that were intractable a decade ago. The confluence of mature formal methods, powerful verification tools demonstrated by HARDENS, urgent safety imperatives documented by persistent human error statistics, and fundamental gaps in current hybrid dynamics treatment creates a compelling and timely research opportunity.

3 Research Approach

This research will overcome the limitations of current practice to build high-assurance hybrid control systems for critical infrastructure. Hybrid systems combine continuous dynamics (flows) with discrete transitions (jumps), which can be formally expressed as:

$$\dot{x}(t) = f(x(t), q(t), u(t)) \tag{1}$$

$$q(k+1) = \nu(x(k), q(k), u(k)) \tag{2}$$

Here, $f(\cdot)$ defines the continuous dynamics while $\nu(\cdot)$ governs discrete transitions. The continuous states x, discrete state q, and control input u interact to produce hybrid behavior. The discrete state q defines which continuous dynamics mode is currently active. Our focus centers on continuous autonomous hybrid systems, where continuous states remain unchanged during jumps—a property naturally exhibited by physical systems. For example, a nuclear reactor switching from warm-up to load-following control cannot instantaneously change its temperature or control rod

position, but can instantaneously change control laws.

To build these systems with formal correctness guarantees, we must accomplish three main thrusts:

1. Translate operating procedures and requirements into temporal logic formulae
2. Create the discrete half of a hybrid controller using reactive synthesis
3. Develop continuous controllers to operate between modes, and verify their correctness using reachability analysis

The following sections discuss how these thrusts will be accomplished.

3.1 $(Procedures \wedge FRET) \rightarrow TemporalSpecifications$

The motivation behind this work stems from the fact that commercial nuclear power operations remain manually controlled by human operators, despite significant advances in control systems sophistication. The key insight is that procedures performed by human operators are highly prescriptive and well-documented. This suggests that human operators in nuclear power plants may not be entirely necessary given today's available technology.

Written procedures and requirements in nuclear power are sufficiently detailed that we may be able to translate them into logical formulae with minimal effort. If successful, this approach would enable automation of existing procedures without requiring system reengineering. To formalize these procedures, we will use temporal logic, which captures system behaviors through temporal relations. Linear Temporal Logic (LTL) provides four fundamental operators: next (X), eventually (F), globally (G), and until (U). These operators enable precise specification of time-dependent requirements.

Consider a nuclear reactor SCRAM requirement expressed in natural language: *"If a high temperature alarm triggers, control rods must immediately insert and remain inserted until operator reset."* This plain language requirement can be translated into a rigorous logical specification:

$$G(HighTemp \rightarrow X(RodsInserted \wedge (\neg RodsWithdrawn\ U\ OperatorReset))) \tag{3}$$

This specification precisely captures the temporal relationship between the alarm condition, the required response, and the persistence requirement. The global operator G ensures this property holds throughout system operation, while the next operator X enforces immediate response. The until operator U maintains the state constraint until the reset condition occurs.

The most efficient path to accomplish this translation is through NASA's Formal Requirements Elicitation Tool (FRET). FRET employs a specialized requirements language called FRETish that restricts requirements to easily understood components while eliminating ambiguity. FRETish bridges natural language and mathematical specifications through a structured English-like syntax that is automatically translatable to temporal logic.

FRET enforces this structure by requiring all requirements to contain six components:

1. Scope: *What modes does this requirement apply to?*
2. Condition: *Scope plus additional specificity*
3. Component: *What system element does this requirement affect?*
4. Shall
5. Timing: *When does the response occur?*

6. Response: *What action should be taken?*

FRET provides functionality to check the *realizability* of a system. Realizability analysis determines whether written requirements are complete by examining the six structural components. Complete requirements are those that neither conflict with one another nor leave any behavior undefined. Systems that are not realizable from their procedure definitions and design requirements present problems beyond autonomous control implementation. Such systems contain behavioral inconsistencies that represent the physical equivalent of software bugs. Using FRET during autonomous controller development allows us to identify and resolve these errors systematically.

The second category of realizability issues involves undefined behaviors that are typically left to human judgment during control operations. This ambiguity is undesirable for high-assurance systems, since even well-trained humans remain prone to errors. By addressing these specification gaps in FRET during autonomous controller development, we can deliver controllers free from these vulnerabilities.

FRET provides the capability to export requirements in temporal logic format compatible with reactive synthesis tools. This export functionality enables progression to the next step of our approach: synthesizing discrete mode switching behavior from the formalized requirements.

3.2 $(TemporalLogic \wedge ReactiveSynthesis) \rightarrow DiscreteAutomata$

Reactive synthesis is an active research field in computer science focused on generating discrete controllers from temporal logic specifications. The term "reactive" indicates that the system responds to environmental inputs to produce control outputs. These synthesized systems are finite in size, where each node represents a unique discrete state. The connections between nodes, called *state transitions*, specify the conditions under which the discrete controller moves from state to state. This complete mapping of possible states and transitions constitutes a *discrete automaton*. Discrete automata can be represented graphically as a series of nodes that are discrete states, with traces indicating transitions between states. From the automaton graph, it becomes possible to fully describe the dynamics of the discrete system and develop intuitive understanding of system behavior. Hybrid systems naturally exhibit discrete behavior amenable to formal analysis through these finite state representations.

We will employ state-of-the-art reactive synthesis tools, particularly Strix, which has demonstrated superior performance in the Reactive Synthesis Competition (SYNTCOMP) through efficient parity game solving algorithms. Strix translates linear temporal logic specifications into deterministic automata automatically while maximizing generated automata quality. Once constructed, the automaton can be straightforwardly implemented using standard programming control flow constructs. The graphical representation provided by the automaton enables inspection and facilitates communication with controls programmers who may not have formal methods expertise.

We will use discrete automata to represent the switching behavior of our hybrid system. This approach yields an important theoretical guarantee: because the discrete automaton is synthesized entirely through automated tools from design requirements and operating procedures, we can prove that the automaton—and therefore our hybrid switching behavior—is *correct by construction*. Correctness of the switching controller is paramount to this work. Mode switching represents the primary responsibility of human operators in control rooms today. Human operators possess the advantage of real-time judgment—when mistakes occur, they can correct them dynamically with capabilities that extend beyond written procedures. Autonomous control lacks this adaptive

advantage. Instead, we must ensure that autonomous controllers replacing human operators will not make switching errors between continuous modes. By synthesizing controllers from logical specifications with guaranteed correctness, we eliminate the possibility of switching errors.

3.3 $(DiscreteAutomata \wedge ControlTheory \wedge Reachability) \rightarrow ContinuousModes$

While discrete system components will be synthesized with correctness guarantees, they represent only half of the complete system. Autonomous controllers like those we are developing exhibit continuous dynamics within discrete states, as described by $f(\cdot)$ in Equation 1. This section describes how we will develop continuous control modes, verify their correctness, and address the unique verification challenges of hybrid systems.

The approach described for producing discrete automata yields physics-agnostic specifications that represent only half of a complete hybrid autonomous controller. These automata alone cannot define the full behavior of the control systems we aim to construct. The continuous modes will be developed after discrete automaton construction, leveraging the automaton structure and transitions to design multiple smaller, specialized continuous controllers.

The discrete automaton transitions are key to the supervisory behavior of the autonomous controller. These transitions mark decision points for switching between continuous control modes and define their strategic objectives. We will classify three types of high-level continuous controller objectives based on discrete mode transitions:

1. **Stabilizing:** A stabilizing control mode has one primary objective: maintaining the hybrid system within its current discrete mode. This corresponds to steady-state normal operating modes, such as a full-power load-following controller in a nuclear power plant. Stabilizing modes can be identified from discrete automata as nodes with only incoming transitions.
2. **Transitory:** A transitory control mode has the primary goal of transitioning the hybrid system from one discrete state to another. In nuclear applications, this might represent a controlled warm-up procedure. Transitory modes ultimately drive the system toward a stabilizing steady-state mode. These modes may have secondary objectives within a discrete state, such as maintaining specific temperature ramp rates before reaching full-power operation.
3. **Expulsory:** An expulsory mode is a specialized transitory mode with additional safety constraints. Expulsory modes ensure the system is directed to a safe stabilizing mode during failure conditions. For example, if a transitory mode fails to achieve its intended transition, the expulsory mode activates to immediately and irreversibly guide the system toward a globally safe state. A reactor SCRAM exemplifies an expulsory continuous mode: when initiated, it must reliably terminate the nuclear reaction and direct the reactor toward stabilizing decay heat removal.

Building continuous modes after constructing discrete automata enables local controller design focused on satisfying discrete transitions. The primary challenge in hybrid system verification is ensuring global stability across transitions. Current techniques struggle with this problem because dynamic discontinuities complicate verification. This work alleviates these problems by designing continuous controllers specifically with transitions in mind. By decomposing continuous modes according to their required behavior at transition points, we avoid solving trajectories through the entire hybrid system. Instead, we can use local behavior information at transition boundaries. To ensure continuous modes satisfy their requirements, we will employ three main techniques: reachability analysis, assume-guarantee contracts, and barrier certificates.

Reachability Analysis: Reachability analysis computes the reachable set of states for a given input set. While trivial for linear continuous systems, recent advances have extended reachability to complex nonlinear systems. We will use reachability to define continuous state ranges at discrete transition boundaries and verify that requirements are satisfied within continuous modes. Recent advances using neural network approximations of Hamilton-Jacobi equations have demonstrated significant speedups while maintaining safety guarantees for high-dimensional systems, expanding the practical applicability of these methods.

Assume-Guarantee Contracts: Assume-guarantee contracts will be employed when continuous state boundaries are not explicitly defined. For any given mode, the input range for reachability analysis is defined by the output ranges of discrete modes that transition to it. This compositional approach ensures each continuous controller is prepared for its possible input range, enabling subsequent reachability analysis without requiring global system analysis.

Barrier Certificates: Finally, we will use barrier certificates to prove that mode transitions are satisfied. Barrier certificates ensure that continuous modes on either side of a transition behave appropriately. Control barrier functions provide a method to certify safety by establishing differential inequality conditions that guarantee forward invariance of safe sets. For example, a barrier certificate can guarantee that a transitory mode transferring control to a stabilizing mode will always move away from the transition boundary, rather than destabilizing the target stabilizing mode.

Combining these three techniques will enable us to prove that continuous components of our hybrid controller satisfy discrete requirements, and thus, complete system behavior. To demonstrate this methodology, we will develop an autonomous startup controller for a Small Modular Advanced High Temperature Reactor (SmAHTR). SmAHTR represents an ideal test case as a liquid-salt cooled reactor design with well-documented startup procedures that must transition through multiple distinct operational modes: initial cold conditions, controlled heating to operating temperature, approach to criticality, low-power physics testing, and power ascension to full operating capacity. We have already developed a high-fidelity SmAHTR model in Simulink that captures the thermal-hydraulic and neutron kinetics behavior essential for verifying continuous controller performance under realistic plant dynamics. The synthesized hybrid controller will be implemented on an Emerson Ovation control system platform, which is representative of industry-standard control hardware deployed in modern nuclear facilities. The Advanced Reactor Cyber Analysis and Development Environment (ARCADE) suite will serve as the integration layer, managing real-time communication between the Simulink simulation and the Ovation controller. This hardware-in-the-loop configuration enables validation of the controller implementation on actual industrial control equipment interfacing with a realistic reactor simulation, providing assessment of computational performance, real-time execution constraints, and communication latency effects. By demonstrating autonomous startup control on this representative platform, we will establish both the theoretical validity and practical feasibility of the synthesis methodology for deployment in actual small modular reactor systems.

This unified approach addresses a fundamental gap in hybrid system design by bridging formal methods and control theory through a systematic, tool-supported methodology. By translating existing nuclear procedures into temporal logic, synthesizing provably correct discrete switching logic, and developing verified continuous controllers, we create a complete framework for autonomous hybrid control with mathematical guarantees. The result is an autonomous controller that not only replicates human operator decision-making but does so with formal assurance that

switching logic is correct by construction and continuous behavior satisfies safety requirements. This methodology transforms nuclear reactor control from a manually intensive operation requiring constant human oversight into a fully autonomous system with higher reliability than human-operated alternatives. More broadly, this approach establishes a replicable framework for developing high-assurance autonomous controllers in any domain where operating procedures are well-documented and safety is paramount.

3.4 Broader Impacts

Nuclear power presents both a compelling application domain and an urgent economic challenge. Recent interest in powering artificial intelligence infrastructure has renewed focus on small modular reactors (SMRs), particularly for hyperscale datacenters requiring hundreds of megawatts of continuous power. Deploying SMRs at datacenter sites would minimize transmission losses and eliminate emissions from hydrocarbon-based alternatives. However, the economics of nuclear power deployment at this scale demand careful attention to operating costs.

According to the U.S. Energy Information Administration's Annual Energy Outlook 2022, advanced nuclear power entering service in 2027 is projected to cost $88.24 per megawatt-hour [21]. Datacenter electricity demand is projected to reach 1,050 terawatt-hours annually by 2030 [22]. If this demand were supplied by nuclear power, the total annual cost of power generation would exceed $92 billion. Within this figure, operations and maintenance represents a substantial component. The EIA estimates that fixed O&M costs alone account for $16.15 per megawatt-hour, with additional variable O&M costs embedded in fuel and operating expenses [21]. Combined, O&M-related costs represent approximately 23-30% of the total levelized cost of electricity, translating to $21-28 billion annually for projected datacenter demand.

This research directly addresses the multi-billion dollar O&M cost challenge through implementations of high-assurance autonomous control. Current nuclear operations require full control room staffing for each reactor, whether large conventional units or small modular designs. These staffing requirements drive the high O&M costs that make nuclear power economically challenging, particularly for smaller reactor designs where the same staffing overhead must be spread across lower power output. By synthesizing provably correct hybrid controllers from formal specifications, we can automate routine operational sequences that currently require constant human oversight. This enables a fundamental shift from direct operator control to supervisory monitoring, where operators can oversee multiple autonomous reactors rather than manually controlling individual units.

The correct-by-construction methodology is critical for this transition. Traditional automation approaches cannot provide sufficient safety guarantees for nuclear applications, where regulatory requirements and public safety concerns demand the highest levels of assurance. By formally verifying both the discrete mode-switching logic and the continuous control behavior, this research will produce controllers with mathematical proofs of correctness. These guarantees enable automation to safely handle routine operations—such as startup sequences, power level changes, and normal operational transitions—that currently require human operators to follow written procedures. Operators will remain in supervisory roles to handle off-normal conditions and provide authorization for major operational changes, but the routine cognitive burden of procedure execution shifts to provably correct automated systems that are much cheaper to operate.

SMRs represent an ideal deployment target for this technology. Nuclear Regulatory Commission certification requires extensive documentation of control procedures, operational require-

ments, and safety analyses written in structured natural language. As described in our approach, these regulatory documents can be translated into temporal logic specifications using tools like FRET, then synthesized into discrete switching logic using reactive synthesis tools, and finally verified using reachability analysis and barrier certificates for the continuous control modes. The infrastructure of requirements and specifications is already complete as part of the licensing process, creating a direct pathway from existing regulatory documentation to formally verified autonomous controllers.

Beyond reducing operating costs for new reactors, this research will establish a generalizable framework for autonomous control of safety-critical systems. The methodology of translating operational procedures into formal specifications, synthesizing discrete switching logic, and verifying continuous mode behavior applies to any hybrid system with documented operational requirements. Potential applications include chemical process control, aerospace systems, and autonomous transportation, where similar economic and safety considerations favor increased autonomy with provable correctness guarantees. By demonstrating this approach in nuclear power—one of the most regulated and safety-critical domains—this research will establish both the technical feasibility and regulatory pathway for broader adoption across critical infrastructure.

4 Metrics for Success

This research will be measured by advancement through Technology Readiness Levels, progressing from fundamental concepts to validated prototype demonstration. The work begins at TRL 2-3 and aims to reach TRL 5, where system components operate successfully in a relevant laboratory environment. This section explains why TRL advancement provides the most appropriate success metric and defines the specific criteria required to achieve TRL 5.

Technology Readiness Levels provide the ideal success metric because they explicitly measure the gap between academic proof-of-concept and practical deployment. This gap is precisely what this work aims to bridge. Academic metrics like papers published or theorems proved cannot capture practical feasibility. Empirical metrics like simulation accuracy or computational speed cannot demonstrate theoretical rigor. TRLs measure both dimensions simultaneously. Advancing from TRL 3 to TRL 5 requires maintaining theoretical rigor while progressively demonstrating practical feasibility. Formal verification must remain valid as the system moves from individual components to integrated hardware testing.

The nuclear industry requires extremely high assurance before deploying new control technologies. Demonstrating theoretical correctness alone is insufficient for adoption. Conversely, showing empirical performance without formal guarantees fails to meet regulatory requirements. TRLs capture this dual requirement naturally. Each level represents both increased practical maturity and sustained theoretical validity. Furthermore, TRL assessment forces explicit identification of remaining barriers to deployment. The nuclear industry already uses TRLs for technology assessment, making this metric directly relevant to potential adopters. Reaching TRL 5 provides a clear answer to industry questions about feasibility and maturity in a way that academic publications alone cannot.

The work currently exists at TRL 2-3. Formal synthesis and hybrid control verification principles have been established through prior research, placing the fundamental approach at TRL 2. The SmAHTR simulation model and initial procedure analysis place specific components at early TRL 3, where proof of concept has been partially demonstrated for individual elements but not integrated. The target state is TRL 5. Moving from current state to target requires achieving three

intermediate levels, each representing a distinct validation milestone:

TRL 3 ***Critical Function and Proof of Concept*** For this research, TRL 3 means demonstrating that each component of the methodology works in isolation. SmAHTR startup procedures must be translated into temporal logic specifications that pass realizability analysis. A discrete automaton must be synthesized with interpretable structure. At least one continuous controller must be designed with reachability analysis proving that transition requirements are satisfied. Independent review must confirm that specifications match intended procedural behavior. This proves the fundamental approach on a simplified startup sequence.

TRL 4 ***Laboratory Testing of Integrated Components*** For this research, TRL 4 means demonstrating a complete integrated hybrid controller in simulation. All SmAHTR startup procedures must be formalized with a synthesized automaton covering all operational modes. Continuous controllers must exist for all discrete modes. Verification must be complete for all mode transitions using reachability analysis, barrier certificates, and assume-guarantee contracts. The integrated controller must execute complete startup sequences in software simulation with zero safety violations across multiple consecutive runs. This proves that formal correctness guarantees can be maintained throughout system integration.

TRL 5 ***Laboratory Testing in Relevant Environment*** For this research, TRL 5 means demonstrating the verified controller on industrial control hardware through hardware-in-the-loop testing. The discrete automaton must be implemented on the Emerson Ovation control system and verified to match synthesized specifications exactly. Continuous controllers must execute at required rates. The ARCADE interface must establish stable real-time communication between Ovation hardware and SmAHTR simulation. Complete autonomous startup sequences must execute via hardware-in-the-loop across the full operational envelope. The controller must handle off-nominal scenarios to validate that expulsory modes function correctly. For example, simulated sensor failures must trigger appropriate fault detection and mode transitions, and loss of cooling scenarios must activate SCRAM procedures as specified. Graded responses to minor disturbances are outside the scope of this work. Formal verification results must remain valid with discrete behavior matching specifications and continuous trajectories remaining within verified bounds. This proves that the methodology produces verified controllers implementable on industrial hardware.

These levels define progressively more demanding demonstrations. TRL 3 proves individual components work. TRL 4 proves they work together in simulation. TRL 5 proves they work on actual hardware in realistic conditions. Each level builds on the previous while adding new validation requirements.

Progress will be assessed quarterly through collection of specific data comparing actual results against TRL advancement criteria. Specification development status indicates progress toward TRL 3. Synthesis results and verification coverage indicate progress toward TRL 4. Simulation performance metrics and hardware integration milestones indicate progress toward TRL 5. The research plan will be revised only when new data invalidates fundamental assumptions. Unrealizable specifications indicate procedure conflicts requiring refinement or alternative reactor selection. Unverifiable dynamics suggest model simplification or alternative verification methods are needed. Unachievable real-time performance requires controller simplification or hardware upgrades. Any revision will document the invalidating data, the failed assumption, and the modified pathway with adjusted scope.

This research succeeds if it achieves TRL 5 by demonstrating a complete autonomous hybrid

controller with formal correctness guarantees operating on industrial control hardware through hardware-in-the-loop testing in a relevant laboratory environment. This establishes both theoretical validity and practical feasibility, proving that the methodology produces verified controllers and that implementation is achievable with current technology. It provides a clear pathway for nuclear industry adoption and broader application to safety-critical autonomous systems.

References

[1] U.S. Department of Energy. Human performance handbook. Handbook DOE-HDBK-1028-2009, U.S. Department of Energy, 2009.

[2] World Nuclear Association. Safety of nuclear power reactors. `https://www.world-nuclear.org/information-library/safety-and-security/safety-of-plants/safety-of-nuclear-power-reactors.aspx`, 2020.

[3] Y. Wang et al. Analysis of human error in nuclear power plant operations: A systematic review of events from 2007–2020. *Journal of Nuclear Safety*, 2025. Analysis of 190 events at Chinese nuclear power plants.

[4] U.S. Nuclear Regulatory Commission. Operators' licenses. 10 CFR Part 55. Code of Federal Regulations.

[5] John G. Kemeny et al. Report of the president's commission on the accident at three mile island. Technical report, President's Commission on the Accident at Three Mile Island, October 1979.

[6] U.S. Nuclear Regulatory Commission. Guidelines for the preparation of emergency operating procedures. Technical Report NUREG-0899, U.S. Nuclear Regulatory Commission, 1982.

[7] International Atomic Energy Agency. Good practices for cost effective maintenance of nuclear power plants. Technical Report TECDOC-1580, International Atomic Energy Agency, 2007.

[8] Gašper Žerovnik et al. Knowledge transfer challenges in nuclear operations. *Nuclear Engineering and Design*, 2023. Analysis of knowledge transfer from experienced operators.

[9] Y. Jo et al. Automation paradox in nuclear power plant control: Effects on operator situation awareness. *Nuclear Engineering and Technology*, 2021. Empirical study of automation effects on operator performance.

[10] International Atomic Energy Agency. Modern instrumentation and control for nuclear power plants: A guidebook. Technical Report Technical Reports Series No. 387, International Atomic Energy Agency, 2008.

[11] D. Lee et al. Autonomous control of nuclear reactors using long short-term memory networks. *Nuclear Engineering and Technology*, 2019. Demonstration of LSTM-based autonomous control in LOC and SGTR scenarios.

[12] IEEE Working Group. Formal verification challenges for nuclear i&c systems. In *IEEE Conference on Nuclear Power Instrumentation, Control and Human-Machine Interface Technologies*, 2019. Discussion of state space explosion in formal verification.

[13] International Atomic Energy Agency. Human error as root cause in severe nuclear accidents. IAEA Safety Report. Analysis of TMI, Chernobyl, and Fukushima accidents.

[14] Lloyd Dumas. Worker error and safety in nuclear facilities. *Journal of Nuclear Safety*, 1999. Study of incidents at 10 nuclear centers.

[15] D. Gertman et al. The spar-h human reliability analysis method. Technical Report NUREG/CR-6883, U.S. Nuclear Regulatory Commission, 2005.

[16] U.S. Nuclear Regulatory Commission. Cognitive basis for human reliability analysis. Technical Report NUREG-2114, U.S. Nuclear Regulatory Commission, 2016.

[17] J. Rasmussen. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(3):257–266, 1983.

[18] George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956.

[19] James Reason. *Human Error*. Cambridge University Press, 1990.

[20] Joseph Kiniry, Alexander Bakst, Michal Podhradsky, Simon Hansen, and Andrew Bivin. High assurance rigorous digital engineering for nuclear safety (hardens) final technical report. Technical Report ML22326A307, Galois, Inc. / U.S. Nuclear Regulatory Commission, 2022. NRC Contract 31310021C0014.

[21] U.S. Energy Information Administration. Levelized costs of new generation resources in the annual energy outlook 2022. Report, U.S. Energy Information Administration, March 2022. See Table 1b, page 9.

[22] Environmental and Energy Study Institute. Data center energy needs are upending power grids and threatening the climate. Web article, 2024. Accessed: 2025-09-29.