SAND2019-3514C

29th Annual INCOSE international symposium
Orlando, FL, USA
July 20 - 25, 2019

Using Systems Theoretic Perspectives for Risk-Informed Cyber Hazard Analysis in Nuclear Power Plants

Adam D. Williams
Sandia National Laboratories*
P.O. Box 5800
Albuquerque, NM 87185-1371
505-844-6779
adwilli@sandia.gov

Andrew J. Clark
Sandia National Laboratories*
P.O. Box 5800
Albuquerque, NM 87185-0748
505-284-2423
ajclark@sandia.gov

Abstract. Nuclear power plants (NPP) use several approaches and tools when assessing vulnerabilities and hazards, typically employing variations of traditional fault tree analysis (FTA). Yet, the Electric Power Research Institute (EPRI) has sponsored research investigating the efficacy of such traditional risk assessment tools to evaluate hazards for digital instrumentation and control systems. One of the conclusions was that no single methodology is befitting to address the complexities of digital hazards. In response, Sandia National Laboratories developed a digital hazards analysis process building on key tenets of Systems-Theoretic Process Analysis (STPA) and Fault Tree Analysis (FTA). This new methodology—*Hazard and Consequence Analysis for Digital Systems* (HAZCADS)—leverages the benefits of these two techniques to provide a more comprehensive, systems approach to hazard analysis for new and complex digital systems. It is anticipated that HAZCADS can improve the ability of the nuclear industry to overcome the technical and regulatory challenges posed by the increased digitization of nuclear power plant systems—thus providing a risk-informed tool for assessing digital systems.

Introduction

The inclusion of digital instrumentation and control (DI&C) into nuclear power plants (NPP) presents new and unique challenges to probabilistic risk assessments (PRA). Traditional PRAs assess failure modes of process components (e.g., “pump fails to run” or “valve fails closed”) and operator actions but typically neglect the failure modes associated with digital I&C. Digital I&C are accompanied by non-traditional failure modes, such as design errors, software flaws, and

*Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA-0003525. **SAND-PEER REVIEW.**

cyber-attack threats. Due to the prevalence of digital I&C in NPPs and the growing threat of cyber-attacks, it's imperative to understand the effects that cyber-attacks can have on NPPs.

The U.S. Nuclear Regulatory Commission (NRC) requires NPPs to submit a cyber security plan as outlined in the Code of Federal Regulations (CFR) 10 CFR 73.54, which states:

> *The licensee shall protect digital computer and communication systems and networks associated with (i) safety-related and important-to-safety functions;(ii) security functions;(iii) emergency preparedness functions, including offsite communications; and, (iv) support systems and equipment which, if compromised, would adversely impact safety, security, or emergency preparedness functions.*

While this federal requirement mandates NPPs to prepare a cyber security plan, it does not provide guidance on how to execute or develop such a plan. With this lack of guidance, NPPs have taken different approaches to meeting this requirement. Some have attempted to use PRA-based approaches—which have been traditionally used to assess the risk from systems, structures, and components (SSC) in the NPP—to prepare a cyber security plan. The Nuclear Energy Institute has provided several documents intended to assist NPPs with addressing 10 CFR 73.54, including *Cyber Security Plan for Nuclear Power Reactors* (NEI 2010) and *Cyber Security Control Assessments* (NEI 2017). These documents provide a deterministic method for identifying critical digital assets. The large number of probable cyber hazards, however, challenge the efficacy of such deterministic approaches. This suggests that a risk-informed approach is necessary to properly assess the importance of digital I&C to the criteria outlined in 10 CFR 73.54.

In response, this paper first introduces some systems-theoretic criteria by which to evaluate the applicability of traditional hazard analysis techniques for identifying cyber security-related hazards. Next, the results of applying these criteria to two analysis techniques—Systems-Theoretic Process Analysis (STPA) and Fault Tree Analysis (FTA)—are summarized. After summarizing the opportunity for, and possible benefit of, combining these two approaches, an initial framework and prototype cyber hazard analysis method is introduced and demonstrated on a notional main feedwater system for a NPP. Lastly, several insights and conclusions regarding this cyber hazard analysis technique are offered on the benefit of incorporating complex systems engineering perspectives into cyber security approaches.

Systems Engineering Concepts as Evaluation Criteria for Risk-Informed Cyber Security

The challenge of meeting the requirements in 10 CFR 73.54 suggested a need to reframe criteria by which to evaluate traditional hazards analysis techniques. In support of ongoing Electric Power Research Institute (EPRI) research to advance the use of hazards analysis methods to assess cyber vulnerabilities, Sandia National Laboratories (Sandia) developed an evaluation rubric leveraging key systems engineering concepts (EPRI 2016). For example, the systems engineering concept of emergent systems behaviors provides an organizing principle for better characterizing NPP operations as resulting from analog process components, digital I&C, and operator actions. Therefore, analytical approaches to support 10 CFR 73.54 must be capable of capturing the impacts of cyber hazards *and* the impacts of the interactions between digital components, analog components, and

operator actions. Accounting for the importance of these interdependencies between digital, physical, and human components within a NPP is another key evaluation criteria leveraged from systems engineering. The complex interrelationship of these different components presents unique challenges to maintaining the as-built/as-designed control of the NPP. Proper assessment of the interrelationship of components requires that control and feedback be accurately modeled and captured in NPPs.

What resulted was a set of evaluation criteria used to assess the ability of different analytical techniques to provide risk-informed cyber security plan, which includes the capability to:

- Determine a "holistic" characterization of the NPP;
- Prioritize risk for a "holistic" system characterization;
- Identify new failure modes unique to DI&C components;
- Describe new interactions enabled by DI&C design features;
- Illustrate new system effects from DI&C-related failure modes and interactions; and,
- Visualize the interrelationships between DI&C and non-DI&C system elements.

In order to not unnecessarily reinvent the wheel, the EPRI-sponsored research applied these criteria to a suite of traditional hazard analysis techniques (EPRI 2015(a)). Here, the aim was to identify a systematic method of characterizing the system, identifying non-traditional failure modes that compromise the intended system control, and provides a means for consequence analysis and risk prioritization. Two current hazard analysis techniques—System-Theoretic Process Analysis (STPA) and Fault Tree Analysis (FTA)—individually measured well against these criteria. Yet, their potential *combination* was further explored and demonstrated an even higher capability to meet the criteria necessary for evaluating risk-informed cyber security at NPPs.

Systems Theoretic Process Analysis (STPA)

Unlike traditional hazard analyses which rely on commonly accepted chain-of-event models, Systems Theoretic Process Analysis (STPA) is a hazard analysis method that uses key tenets of systems and control theory for explaining undesired system behaviors (Leveson 2011). STPA consists of four main steps (Leveson & Thomas 2018): define the purpose of the analysis, model the hierarchical control structure (HCS), identify unsafe control actions, and identify loss scenarios. The HCS is a system model that explicitly includes the communication between controllers and a controlled process. STPA uses control actions and feedback signals to illustrate the communication between controllers (whether physical, digital, or human) and a controlled process (e.g., normal NPP operations). The control actions are a function of the process models, control algorithms, and feedback signals that are built into the components and systems. STPA's basis in systems and control theory enables this technique to identify nontraditional causal scenarios and hazardous control actions that can lead to an unacceptable system loss.

This process further asserts that system losses result from flawed interactions between physical components, engineering activities, operational mission, organizational structures and social factors (Leveson 2011). STPA can identify hazardous system states that result when a malfunction in one component leads to a system loss *and* when all components behave in an expected manner, but the system still experiences a loss. For example, a digital controller may be programmed to shut down the system when a pressure threshold is exceeded in a reactor. However, during startup the operators anticipate that the pressure threshold will be exceeded, but this pressure increase is

considered operationally acceptable because the NPP is in a startup mode. In this scenario, the controller would always override the operator's intent and shutdown the reactor during startup. Although the components operate as expected, the observed (or, *emergent*) behavior still results in a system loss. These types of hazards are typically missed by other analysis methods when accident scenarios for evaluation are identified *a priori*. STPA analysis models may uncover design errors, software flaws, component interaction accidents, and cognitively complex human decision-making errors (Leveson 2011).

Fault Tree Analysis (FTA)

Fault tree analysis (FTA) is a top-down, deductive failure evaluation technique whereby various chain-of-event pathways to undesired states of the system are identified and prioritized. Once modeled as a fault tree, the system is analyzed by finding all credible event(s) in which the undesired event can occur. FTA is based on the deductive structure where "lower" events (e.g., component failure modes) pass through *AND* and *OR* logic gates, which themselves represent the relationship of events needed for the occurrence of the "higher" events. The lowest level of the fault tree is comprised of primary (or, basic) events that describe specific component failure modes that initiate failure pathways. FTA results in a Boolean equation that includes all the unique combinations of different component failures that can result in the "top" event failure (often representing a system failure). Each term (or combination of faults) is referred to as a *cut set*. Simple fault trees can be solved by hand, but typically software programs are used to perform the Boolean algebra of large and complicated fault trees. FTA has been applied as a method to study system design for over fifty years (NRC 1981).

FTA can be used to quantify the failure probability of a system or collection of systems—or, conversely, estimate their reliability—through the implementation of probabilities of the primary events into the Boolean equations. However, the utility of FTA is not exclusive to quantitative probabilistic results. Fault trees yield qualitative insights regarding the design of a plant and its systems. Cut sets can be assessed qualitatively to determine the defense-in-depth—or lack thereof—of a system. Qualitative insights are important for identifying the strengths and weaknesses of a system.

Hazards and Consequence Analysis for Digital Systems (HAZCADS)

EPRI's evaluation of the STPA and FTA hazard analysis techniques concluded that neither of them individually could identify, evaluate, and prioritize cyber security hazards pursuant to the criteria for an all-inclusive hazard analysis technique. Yet, additional discussions indicated a potential utility for combining these two methods. Follow-on analysis indicated that this particular combination of methods best leveraged the analytical benefits of each respective technique. For example, FTA struggles to identify new failure modes unique to DI&C components, whereas STPA excels at this criterion. The final conclusion of this hazards analysis evaluation was that combining STPA and FTA leverages the benefits and overcomes the shortcomings of the individual methodologies to meet the criteria for risk-informed cyber security methodology that can meet the criteria outlined in 10 CFR 73.54.

The result of this EPRI-sponsored research was the development of the *Hazards and Consequences Analysis for Digital Systems* (HAZCADS) analysis technique. HAZCADS merges the principles and elements of STPA and FTA to efficiently and methodically address hazards and

consequences that can emerge from digital systems. HAZCADS uses specific steps and outcomes from each technique without significantly altering either STPA or FTA. HAZCADS Step 1 (a step essential to all hazard analysis methods) is to gather plant design, system design, and hazard data. Although the numeric labeling of HAZCADS Steps 2-4 implies sequential processes, HAZCADS Steps 2 and 3 (STPA Steps 1-3) and HAZCADS Step 4 (FTA Steps 1-3) may be done in parallel. HAZCADS steps 5 and 6 provide a rigorous and systematic process for evaluating all potential cyber hazards. The HAZCADS analysis technique is illustrated in Figure .

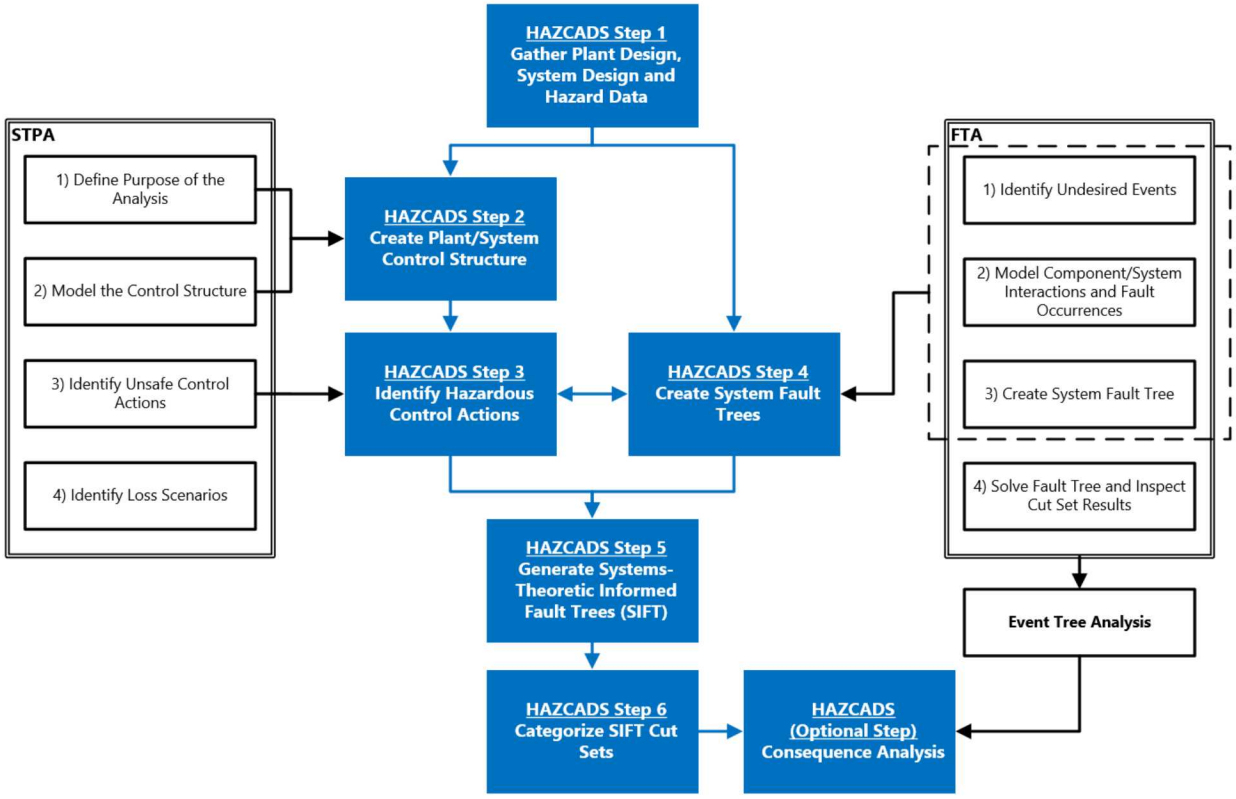

Figure 1. Graphical representation of HAZCADS (blue) that identifies where analytical aspects of STPA and FTA are incorporated.

More specifically, in steps 2 and 3, HAZCADS leverages STPA's hierarchical control structure (HCS) and hazardous control actions. The HCS, as a complex systems model, has two benefits. First, the HCS merges piping and instrumentation diagrams (P&ID) and digital network topologies. Typically, these two mappings are created independently and evaluated from different perspectives. By merging the P&ID and digital network topology, a more complete mapping of the system is provided in a single diagram. Second, the HCS models control actions and feedback between controllers (whether physical, digital, or human) and a controlled process (e.g., normal NPP operations). These control actions and feedbacks identify how control—or the lack thereof—propagates through the system. The hazardous control actions represent potential digital I&C failure modes that can be incorporated into fault tree models.

Similarly, the primary outputs from FTA used in HAZCADS are the fault tree models. The fault tree models describe the relationship between process components that are required to perform the intended function of the system and illustrate specific combinations of components that lead to a

system loss. HAZCADS inserts the hazardous control actions generated using the early steps of STPA as undeveloped events into traditional fault tree models. Modeling STPA hazardous control actions as undeveloped events give awareness to the fact that hazardous control actions still need to be fully comprehended (e.g., through causal analysis).

Incorporating hazardous control actions into fault tree models leads to a fundamentally new model called "systems-theoretic informed fault trees," or SIFTs. SIFTs better incorporate both the direct and indirect roles of digital components in potential failure pathways. Where traditional FTA generally employs probabilities to quantitatively describe cut sets, the HAZCADS approach does not attempt to quantify the SIFT cut sets. Rather, these new fault tree models are solved using the same Boolean algebraic logic that has been the foundation for FTA for decades, with the resultant cut sets evaluated qualitatively. HAZCADS Step 6 uses the *operational* and *state-of-being* (e.g., physical versus digital) differences between components included in identified failure pathways to categorize the cut sets. HAZCADS Steps 1-6[1] can be applied across all the safety and non-safety systems in a nuclear power plant and will result in cut sets describing a range of potential failure pathways across the complex system.

An Illustrative Example: HAZCADS Analysis of a Notional Main Feedwater Control System

HAZCADS was applied to a notional main feedwater control system (MFWCS) for a nuclear power plant. Typically, the MFWCS's primary function is to remove heat from the reactor coolant system (RCS). The MFWCS is the preferred mechanism for removing heat from the RCS because it transfers thermal energy through the steam generators in a manner less intrusive to normal operations. The MFWCS is designed such that a single steam generator can remove sufficient heat to protect the nuclear reactor. Plant operators prefer to use the MFWCS system if it is available, but an auxiliary feedwater system exists as a backup. Although this notional MFWCS example does not reference of any specific nuclear power plant, the piping, instrumentation, and digital network topology for the MFWCS system are representative of commercial NPPs.

For this illustrative example, the HCS is overlaid onto a P&ID for a notional MFWCS as shown in Figure 2. In this example, the hazard of concern is the loss of the MFWCS which results from a loss of steam generator 1 AND steam generator 2.

---

[1] Figure 1 identifies an *optional* step in HAZCADS—consequence analysis. If SIFTs are created for all digital systems (or physical systems with digital components), the SIFTs can easily be integrated into an event tree model. Event tree-related analysis techniques can be used for consequence assessment to describe the (potential) contributions of digital I&C. This is a further area of current R&D.

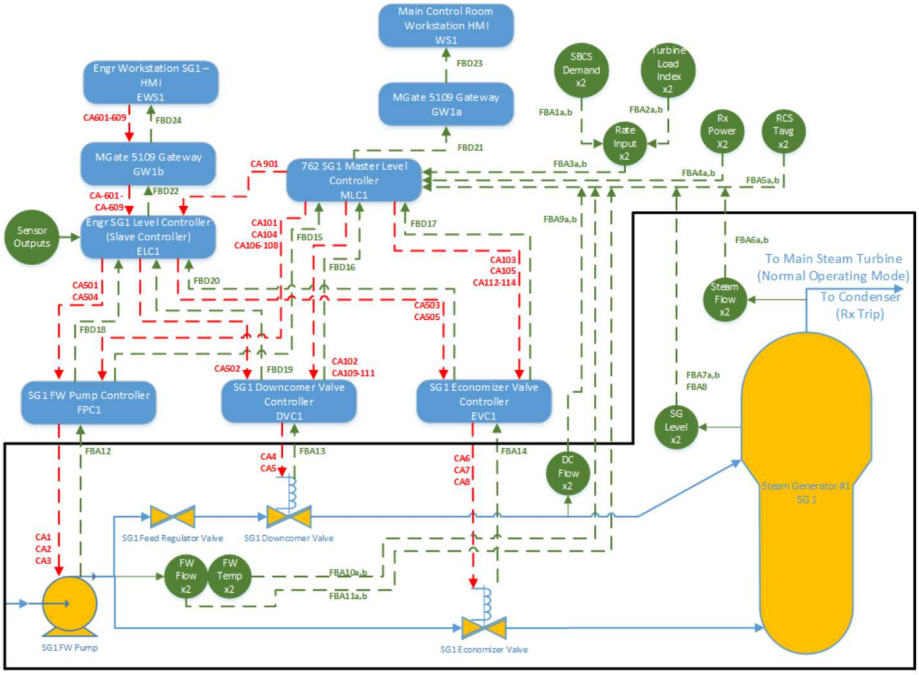

Figure 2. Notional MFWCS digital and physical systems modeled as an STPA-related hierarchical control structure.

The creation of the SIFT for the MFWCS system significantly expanded the size of the fault tree model, which in turn generated an increased number of cut sets for the loss of the MFWCS. More specifically, whereas traditional FTA identified *36* cut sets, evaluation of the SIFT generated *256* cut sets. Incorporation of digital components into the fault tree models led to the realization of SIFT cut sets that can be differentiated into three specific categories:

- **Type 1**: Cut sets comprised solely of *non-digital hardware* component failures;
- **Type 2**: Cut sets comprised of *combinations* of hazardous control actions with non-digital hardware component failures; and,
- **Type 3**: Cut sets comprised only of *hazardous control actions*.

Table 1. Summary of HAZCADS results for evaluating a notional MFWCS.

SIFT Cut Set Category	Number of Cut Sets	% of Generated Cut Sets
Type I	36	14
Type II	120	47
Type III	100	39

As illustrated in Table 1, the 256 SIFT cut sets consist of qualitatively different pathways to a loss. Type I cut sets should be *the same* as cut sets identified by traditional FTA alone. This is validated as the 36 Type I SIFT cut sets match those identified in evaluating the traditional fault tree. These SIFT cut sets provide two unique capabilities. First, Type 3 cut sets identify potential digitally related faults (or cyber exploitation vulnerabilities) of the DI&C system that result in hazards, as well as reveal control action violations from a single digital component. Second, if hazardous

control actions associated with specific digital assets occur only in Type 2 cut sets (e.g., combinations of digital events and non-digital events), then those digital assets represent new opportunities for potential attacks on the NPP—and, therefore, also identify where mitigation measures might need to be implemented. In addition to identifying the digital I&C components that contribute to the hazard, HAZCADS can also identify digital I&C components that *do not* contribute to the loss of MFWCS hazard. Comparing the digital I&C components built into the system to the components that occur in a cut set, it was found that eight of 18 total digital components in the system do not contribute to the MFWCS loss—the implication of this finding is that limited cyber security resources can be focused on more vulnerable areas.

Conclusions, Insights, & Next Steps

Rather than focusing on prioritizing hazards *a priori* (which tends to limit the analytical scope of other techniques), the rigorous and systematic process of identifying hazards in HAZCADS offers several unique characteristics. First, HAZCADS favors traceability—the ability to provide a clear, logical path from high-level system objectives to individual FTA causal paths and back—to explain a range of hazards with different potential (and difficult to equivalently measure) consequences. Combining STPA and FTA helps to both identify non-traditional failure modes but also helps trace their pathways via fault tree models. For example, the MFWCS example identifies 10 of 18 components that are important in preventing the loss of MFWCS hazard *and* provides a traceability from the pathway back to the HCS. As such, the SIFT cut sets show that system hazards can be achieved *entirely* from digital component interactions; namely, hazardous control actions related to digital assets. Furthermore, the HAZCADS analysis identified 8 of 18 digital components that *do not* have any impact on the loss of MFWCS hazard.

Second, HAZCADS favors flexibility—the ability to scale from evaluating a small list of hazards to an expanded, and increasingly complex, list of hazards—over being able to formally optimize analytical results for a select set of undesired losses. The evaluation framework within HAZCADS could be applied to supply chain threat identification or cyber attack vulnerability mitigation. As one of the powerful results of HAZCADS, it is designed to be "domain agnostic." For example, because HAZCADS builds on FTA, other concepts that similarly build on FTA—like digital common-cause failures (CCF) and defense-in-depth (DID)–can also be incorporated into this HAZCADS framework. Considering that FTA has been applied to vital area identification for security of nuclear facilities (Varnado and Whitehead 2008), HAZCADS could also be applied to physical security related hazards of digital I&C.

Despite these benefits, it is important to note that HAZCADS does not include the causal analysis approach in STPA Step 4. As described above, it is the failure *modes*, not the failure *mechanisms*, that are incorporated into fault trees. In this regard, the causal analysis is not necessary for incorporating non-traditional digital failure modes into fault trees. Although failure mechanism assessment results are implicitly included in basic event analysis, understanding all the failure mechanisms is not necessary for performing a risk-informed analysis.

From the perspectives of meeting 10 CFR 73.54, HAZCADS can possibly be used to develop a NPP cyber security plan that also provides a risk-informed method that provides a rigorous, systematic process for evaluating cyber hazards that surpasses current approaches. In addition to supporting the development of higher fidelity cyber security plans, HAZCADS is also able to

identify digital components which have no impact on system hazards and can help streamline cyber security mitigations. Overall, HAZCADS offers higher fidelity assessment, more flexible vulnerability identification, and a hazard prioritization schema capable of identifying cyber-related hazards in nuclear power plants, in particular, and in complex systems with digital components, more broadly.

References

Electric Power Research Institute (EPRI), 2015(a), *Analysis of Hazard Models for Cyber Security – Phase I*, Palo Alto, CA.

Electric Power Research Institute (EPRI), 2015(b), *Program on Technology Innovation: Cyber Hazards Analysis Risk Methodology – Phase II: A Risk Informed Approach*, Palo Alto, CA.

Electric Power Research Institute (EPRI), 2016, *Cyber Security Technical Assessment Methodology – Vulnerability Identification and Mitigation*, Palo Alto, CA.

Leveson, N. and J. Thomas, 2018, *STPA Handbook,* Partnership for Systems Approaches to Safety and Security (PSASS), < http://psas.scripts.mit.edu/home/materials/>

Leveson, N., 2011, *Engineering a Safer World – Systems Thinking Applied to Safety,* MIT Press, Cambridge, MA.

Nuclear Energy Institute (NEI), 2010, *Cyber Security Plan for Nuclear Power Reactors*, NEI 08-09.

Nuclear Energy Institute (NEI), 2017, *Cyber Security Control Assessments*, NEI 13-10, Revision 5.

U.S. Code of Federal Regulations (CFR), 2010, *10 CFR 73.54: Protection of digital computer and communication systems and networks*, <https://www.gpo.gov/fdsys/granule/CFR-2012-title10-vol2/CFR-2012-title10-vol2-sec73-54>

U.S. Nuclear Regulatory Commission (NRC), 1981, *Fault Tree Handbook*, NUREG-0492.

Varnado G. and D. Whitehead, 2008, *Vital Area Identification for U.S. Nuclear Regulatory Commission Nuclear Power Reactor Licensees and New Reactor Applicants*, SAND2008-5644, Sandia National Laboratories, Albuquerque, NM.

Biography

Adam D. Williams. As a Senior R&D System Engineer in the Center for Global Security and Cooperation at Sandia National Laboratories, Dr. Williams he serves as program manager for all work supporting the Department of State's Partnerships for Nuclear Threat Reduction (DOS/PNTR). He is also a systems-theoretic analysis expert supporting projects in managing complex risk, system dynamics, physical protection system development and analysis, and global engagement, including for Laboratory Directed R&D (LDRD), Electric Power Research Institute (EPRI) and National Nuclear Security Administration (NNSA) initiatives. He earned his PhD in the Engineering Systems: Human-Systems Engineering from the Massachusetts Institute of Technology in 2018.

Andrew J. Clark. As a R&D Nuclear Engineer in the Center for Energy and Earth Systems at Sandia National Laboratories, Mr. Clark is a hazard, safety, and risk analyst for high-consequence systems. Andrew has spent his career performing probabilistic risk assessments, primarily for nuclear power plants. Andrew has performed Level 1-3 PRA for nuclear power plants and has experience using PRA and thermal-hydraulic computer codes for nuclear accident analysis. His recent research has focused on applying PRA tools to U.S. weapon systems and digital instrumentation and control systems. Andrew received his M.S. in nuclear engineering from The Ohio State University in 2015.