Applying 3S Lessons: Using Safety

Concepts to Develop

"Risk-Informed Safeguards" for

Small Modular Reactors

C.A. FAUCETT

Sandia National Laboratories

Albuquerque, NM, United States of America

Email: cfaucet@sandia.gov

A.D. WILLIAMS

Sandia National Laboratories

Albuquerque, NM, United States of America

**Abstract**

The paper introduces an alternate perspective to integrating security/safety/safeguards where concepts from nuclear safety are explored to better communicate risks in the international safeguards domain. Recently, the safety concept of risk significance (defined as a level of risk exceeding a predetermined threshold) and organizational gradients was used to develop a conceptual “insider risk significance” framework for multi-insider threats. Using key similarities between State-level diversion and subnational theft, this paper conceptualizes a similar “safeguards risk significance” framework for better characterizing safeguards risks for small modular reactor deployment. While development of this framework is preliminary, the concepts presented have potential to enhance communication of safeguards risks to security and safety experts.

## Introduction

As evidenced by multiple evaluation approaches, the risks associated with the failure to safeguard nuclear technology—sometimes referred to as “proliferation risk”—is a deeply studied, well-understood phenomenon. Yet, in safety, security, and safeguards literature, safeguards is often perceived as the least important of the 3S domains despite its potential for extraordinarily high consequences if neglected. This is not to say that safeguards is unappreciated by non-subject matter experts. Rather, because of natural parallels between safety and security (e.g., similar perception of consequences and owner/operator responsibilities) safeguards may not receive adequate attention in a cross-domain context. Developing clearer descriptions of the potential for incredibly high consequence events related to safeguards and better characterizing the importance of safeguards-by-design can help support the responsible development of advanced or small modular reactors.

To this end, this paper presents an alternate perspective of safety, security, and safeguards that focuses more specifically on how concepts or terminology from one domain can be used to facilitate a common understanding of risk between the other domains. Specifically, this paper will discuss how the concept of “risk significance,” which is commonly used in nuclear safety but recently has been applied to characterize security risks from insider threats, could be used to enhance communication of safeguards or proliferation risk. The objective of this paper is not to present a completed safeguards risk significance model. Instead, this paper explores how a concept like risk significance can be used to more clearly and completely express safeguards risk to security, safety, or other non-safeguards experts.

This paper is organized into a single primary section where the concept of risk significance is introduced, and an overview of how risk significance was used to model insider threats at nuclear facilities is presented. Then, discussion is provided about how risk significance could be similarly be applied to the safeguards domains and what some of the challenges in doing so may be. After this section, a brief conclusion is provided.

## Safeguards Risk Significance

### Risk and Risk Significance

A commonly-accepted definition of risk is that risk is a triplet of three questions [1]:

* What can go wrong?
* How likely is it?
* What the consequences if it does?

The information provided by answering this triplet of questions can be used by decision makers to establish whether a system is “good enough” to mitigate risks of concern as defined by the various decision makers or governing regulatory bodies. For example, determining whether a system is “safe enough” requires a threshold. For nuclear safety in the U.S., the Nuclear Regulatory Commission (NRC) uses the concept of risk significance to identify “safe enough.” More specifically, the NRC defines risk significant as either a system, structure, component, or accident sequence that exceeds a predetermined risk contribution level for a facility or a level of risk that exceeds an established level of significance [2].

Similarly, a publication by the Nuclear Energy Institute (NEI)—titled “Risk-Informed Performance-Based Technology Inclusive Guidance for Non-Light Water Reactor Licensing Basis Development” (NEI 18-04)—uses the concept of risk significance to establish design objectives for nuclear safety [3]. Summarized briefly, NEI 18-04 provides guidance for risk-informed, technology neutral, and performance-based reactor licensing in terms of a two-dimensional risk space. This risk space, illustrated in Figure 1 (recreated in [4] from [2]), is bounded by axes for event frequency in units of yr-1 and consequences with units of dose. In this risk space, quantitative anchors for regulatory event frequencies and dose limits are used to create a frequency-consequence (F-C) target line. NEI 18-04 then defines risk-significant licensing basis events as events with frequencies within 1% of the F-C target line with site boundary doses exceeding 2.5 mrem. The F-C line is not a cutoff, but rather an objective for nuclear facility design i.e., reactor systems should be designed such that licensing events are below this F-C line and therefore not risk significant.



Figure 1. Illustration of Frequency-Consequence curve (recreated in [4] from [2]).

### Insider Security Risk Significance Model

Although the risk significance concept is rooted in nuclear safety, recent nuclear security research found that risk significance had utility for explaining multi-insider security threat risks. In studying why certain combinations of personnel present greater potential for undesirable consequences as multi-insider threats, it was found that risk significance can be borrowed from nuclear safety to help more clearly and comprehensively characterize insider threat mitigation efforts [4].

Drawing on these parallels between radiological consequences in safety and security, Faucett defined *risk significant insider threat* in terms of an insider whose capabilities to create undesirable consequences exceeded the relevant security measures at a facility (e.g., access control, separation of duties) [4]. Like safety, where the design objective is to move events away from a risk significance boundary, the objective of insider threat mitigation is framed as applying security measures such that all facility personnel are far removed from *insider* risk significance. In this way, risk significance provided an intuitive way to describe insider threat mitigation initiatives.

However, unlike nuclear safety where historical data and performance testing can be used to quantify risk metrics, including frequency and dose, there is virtually no data for nuclear insider threats from which a quantitative risk significance threshold can be defined (let alone insider attacks involving more than one insider). Additionally, quantitative objectives are difficult to define for insider threat mitigation. Unlike nuclear safety, even “minor” or “low” consequence insider threat events (i.e., analogous to anticipated operational occurrences) are generally not considered acceptable. This indicates that the only obvious quantitative objective is zero unmitigated insider threats. Because of this, creating a quantitative threshold for insider security analogous to an F-C curve is difficult.

Because of the issues related to establishing an insider risk significance threshold, elements of Rasmussen’s abstraction of organizational safety were incorporated for an alternate interpretation of risk within this insider risk significance model [5]. Rasmussen describes work activities as existing within a space of possibilities bounded by limits of functionally accepted performance, a recreation of which is shown in Figure 2. Over time, organizational dynamics (namely the internal human tendency to achieve desired work objectives with the least amount of effort possible or existence of management pressures for increased efficiencies) can change how these work activities are achieved. Rasmussen further postulates that such dynamics create natural gradients in the Brownian motion of an organization that draw work towards the acceptable performance boundary and can ultimately result in a safety accident when the boundary is crossed. Risk management, therefore, requires *counter* gradients in the form of safety awareness campaigns or enhanced safety culture. In terms of insider risk significance, a core component of Rasmussen’s abstraction is that movement towards or from the boundary of acceptable performance is more important than the boundary itself. Likewise, it was postulated in [4] that the threshold for insider risk significance is *less* important if the “gradients” that move an insider towards risk significance are known.



Figure 2. Representation of Rasmussen's dynamic safety model (recreated from [5]).

The final multi-insider risk model consisted of a two-dimensional space with an “insider risk significance” threshold not like Rasmussen’s acceptable performance boundary. Rather than a boundary of acceptable performance, the boundary represented a space in which an insider’s or insider pair’s capabilities could overcome preventative insider mitigation measures and allow them to successfully execute an attack i.e., a “risk significant” insider. By abstracting insider risk factors from literature, the final insider risk significance model considered two axes representing an insider’s ability to avoid detection prior to and after malicious actions had occurred, respectively. Although the conceptual insider risk model could not be validated due to a lack of nuclear insider data, the model generated new insights into insider risk that helped explain findings from an artificial intelligence-based insider detection study which demonstrates the utility of such conceptual models.

### Potential for Safeguards Risk Significance

Thus far, this paper has focused on using safety concepts to describe security risks from insider threats. However, besides the obvious connection regarding unauthorized, covert movement of nuclear material, there are conceptual similarities between the insider theft and State-level diversion that suggest potential for a “safeguards risk significance” model.

Safeguards risks may be best described using a gradient concept as was done for insider risk significance. As postulated by Rasmussen, there is benefit in using gradients to describe safeguards as gradients—facilitating analysis of why safeguards events occur and attribution mitigation effectiveness leveraged against the safeguards risk significance threshold. That is, treating risk as gradients towards or away from unacceptability can shift focus away from go/no-go performance boundaries such as whether proliferation has occurred.

Like insider threats, safeguards risk is effectively about potential for misuse. Insider risk is unavoidable in that personnel must be granted special access, authority, or knowledge privileges for an organization to function. Yet, it is precisely these privileges that enable insiders to perform malicious actions if they are motivated to do so. Similarly, signatories of the Treaty on the Nonproliferation of Nuclear Weapons (NPT) have an inalienable right to the peaceful use of nuclear energy under Article IV [6]. Yet, peaceful nuclear technology can easily be repurposed for non-peaceful purposes. In this way, insider and safeguards risk share a commonality in that the risk is primarily driven by “what can go wrong?” versus “how likely is it?”.

Finally, in contrast to safety where quantitative approaches can be used to define acceptability thresholds in a technically defensible manner, both insider security and safeguard risk evaluations tend towards subjectivity for defining such thresholds. Here, subjectivity is used to contrast with safety where probabilistic risk quantification approaches can be established based on real world performance and testing data. This is not to say that the factors that contribute to insider and safeguards risks are not well studied. Consider, for example, how risks associated with an insider threat can be described using the insider’s access, authority, and knowledge or how non-proliferation risks can be expressed through metrics such as material attractiveness, reactor attributes such as neutron energy spectrum, or the safeguards agreements the owner/operator State has with the IAEA.

### Challenges with Safeguards Risk Significance Concept

Although commonalities between insider theft and State-level diversion suggest potential for a safeguards risk significance model, there are certain attributes of safeguards risk that challenge the appropriateness of the risk significance concept. These challenges, amongst others, will need to be considered should a safeguards risk significance model be further developed.

One challenge in developing a safeguards risk significance model is how to verify and validate a target risk threshold such as an F-C curve to compare whether nuclear technology is safeguards risk significant. The NPT establishes a universally agreed baseline for acceptable safeguards risk in that, again, all signatories are entitled to the peaceful use of nuclear energy if the nuclear material is placed under safeguards. Appropriating safety parlance, this threshold would effectively need to answer “how much is safeguarded enough?”. Given that the IAEA uses a risk-informed State-Level Concept, establishing a universal threshold is not expected to be straightforward.

A second barrier to safeguards risk significance that safeguards/non-proliferation risk is a complex blend of fundamental nuclear physics, State technological capabilities, safeguards agreements between States and the IAEA, and interregional/international geopolitics. This suggests that describing safeguards risk comprehensively may require more granularity or multidisciplinary coordination than the probabilistic or organizational gradient analogues from other risk significance models can provide.

Finally, and unique to safeguards, a State may exercise its sovereignty and withdraw from the NPT under Article X.1, terminating its safeguards agreements with the IAEA [6]. Considering a State’s right to withdraw from the NPT complicates defining an unacceptable risk threshold as it could be argued that any transfer of nuclear power technology capable of producing weapons useable nuclear material is “risk significant”. Furthermore, there is no analogue in the security or safeguards domains that could be leveraged to represent this risk i.e., an insider cannot declare security measures no longer apply and freely commit malicious actions. It is therefore presently unknown how this risk could be represented in a safeguards risk significance model or what frameworks would be required to do so.

## Conclusion

This paper presented high-level discussion regarding how well the safety concept of risk significance could be applied to model or describe safeguards risks. Risk significance proved useful in characterizing risks associated with insider threats at nuclear facilities, and there are similarities between insider threats and State-level diversion that suggest potential for using risk significance to present safeguards risks including the potential utility of a gradient-based approach to risk and challenges in establishing a quantitative risk significance threshold.

However, the safeguards domain faces unique challenges not seen in safety or security (notably the ability of States to withdraw from safeguards agreements) that complicate the use of risk significance for describing safeguards risks, notably the complexity of safeguards risk evaluations.

If successfully develop, a safeguards risk significance model could help communicate or explain safeguards risks to non-safeguards experts and facilitate more robust security-, safeguards-, and safety-by-design approaches for advanced reactors. To realize these benefits, the next steps in developing a fully realized safeguards risk significance model include structuring the abstract, conceptual model described in this paper into systems model that maps the relationships between safeguards risk factors and safeguards “risk significance”.

ACKNOWLEDGEMENTS

The authors gratefully thank the National Nuclear Security Administration’s Advanced Reactor International Safeguards Engagements program for supporting authorship of this paper. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

1. KAPLAN, S, GARRICK, B.J., On The Quantitative Definition of Risk, Risk Analysis **1** 1 (1981), 11-27.
2. Risk-Informed Performance-Based Technology Inclusive Guidance for Non-Light Water Reactor Licensing Basis Development, Nuclear Energy Institute (2019).
3. United States Nuclear Regulatory Commission, Risk significant (2023), <https://www.nrc.gov/reading-rm/basic-ref/glossary/risk-significant.html>
4. FAUCETT, C., Development of a Conceptual Multi-Insider Risk Model for Nuclear Facilities, Texas A&M University, College Station, TX (n.p.)
5. RASMUSSEN, J., Risk management in a dynamic society: a modelling problem, Safety Science **27** 2-3 (1997), 183-213.
6. Treaty on the Non-Proliferation of Nuclear Weapons, INFCIRC/140, IAEA, Vienna (1970).