A New Safety Paradigm for Successful Next Generation Nuclear Systems Integrating AI and Automation – For Sustaining the Nuclear Leadership in Safety Realization

Yong Hee Lee
Advanced I&C Research Division, *Korea Atomic Energy Research Institute*
Daedeok-daero 989-111, Daejon, Korea, 34050
**Corresponding author:: yhlee@kaeri.re.kr, yhlee0412@gmail.com*

***Keywords :** positive safety, AI and Automation, safety paradigm, resilience, Safety III*

1. Introduction

Nuclear power has been realized before man-kind as an energy for electric power generation as well as bombs. Safety was the prerequisite and the top priority to actually show the positive paradigm of *'Atom as a Worker not a Soldier'* in early era. This paper reexamines the scope of safety that nuclear systems have led and achieved so far and reviews inadequate or future considerations for better safety. In particular, five safety paradigms were presented that are necessary to introduce AI and automation into nuclear systems as key technologies for the future. A new paradigm and five starting points proposed to deal with safety in next-generation nuclear systems such as SMR.

2. A Brief Review on the Safety Developments in Nuclear

2.1 Safety Achieved and Remaining in Nuclear

Safety in nuclear power is a pre-requisite for nuclear technology, so more effort has been made than nuclear technology itself. As a full-fledged technology, safety began in the chemical and military fields, but nuclear power has taken the traditional safety technology to a new level, regardless of the deterministic and probabilistic approaches. This can be attributed to the continuous realization of safety through engineering technology development according to the technical and safety characteristics of nuclear power that can be summarized as follows. (2018/2019 Lee)

- technological aspects : large-scale, complex, tightly-coupled, public energy, etc.
- safety aspects : rare-event, unfamiliar, irreversible, non-injury, catastrophic, etc.

Korea, which has recently emerged as a new leader in nuclear technology, is striving to satisfy the various safety perspectives required as well as the new level of demand. Today, nuclear safety has rapidly expanded its conceptual scope to cover various dimensions, including capital and social risks, as well as future risks such as climate/environmental destruction and generational/genetic problems, beyond basic risks such as first fatality and injury and loss. (2023 Lee)

- Scope of safety spread: quality safety, parts safety, functional safety, system safety, life safety, human-technology-organization (HTO) safety, social safety, environmental safety, genetic and future generation safety

Despite its consistently pioneering efforts, nuclear safety has recently faced the challenge of two new big waves to make matters worse.

First, it is a rapid increase in the realistic demand for safety. With the development of society, expectations and demands for safety are rapidly changing. This is because the development of social networks has not only made it a super-connected society, but also an era of hyper-sensitive responses to safety. This is not only a simple increase in technical standards, but the complex and complex safety requirements act as a great uncertainty in securing safety. Considering the publicity of nuclear power as an energy that provides the basis of society, the surge in social demands for safety poses a serious challenge.

Second, it is the spread of the practical scope of safety. With the advancement of technology, the reality of safety, in which various functions and their losses are possible at the same time, is changing. As computer and software-based technologies lead, many functions that were difficult to implement in the past provide various benefits, but at the same time, the possibility of losing these new values and benefits becomes a safety challenge as it is. This is because technological advances also act as vulnerability. Advanced technology is mass-producing functions that are not sufficient for proper pre-verification and even experience due to rapid changes, so traditional safety technology is controversial not only in nuclear safety but also in allowing commercial services.

2.2 Safety Challenges with AI and automation

These two waves that have come to safety have emerged as more serious challenges in relation to AI and automation technology. Basically, the introduction of AI and automation is positively expected to have a great effect in practically handling the technical and safety characteristics of nuclear systems. The role of responding to very low-frequency defects and

uncertainties arising from large-scale complex and technically-coupled systems has been a mission given to humans and organizations. As seen in the TMI, Chernobyl, and Fukushima accidents, it was difficult to say that the reliability of humans and organizations for safety tasks that were sufficiently experienced or unexpected was sufficient to match the ultra-low frequency thinking and high reliability system required for nuclear power. Conceptual expectations for AI and automation technology are inevitable.

AI and automation technologies are expected to provide fundamental developments in resolving the uncertainty of human and human factors. AI and automation technologies, which are rapidly developing in recent years, are expected to fundamentally compensate for human limitations and weaknesses. However, from the experience of new technologies so far, the possibility of new problems cannot be excluded. It can be seen that various AI safety requirements proposed in each country and internationally are pouring out for the safe application of AI. This is because, like all technologies, the technology itself cannot be perfect, and human limitations using the technology remain. Therefore, everyone expects that ethical principles for AI, the three classical robot principles, and the legal requirements for artificial intelligence safety are not enough. The limitations and problems of these AI technologies are experienced already as accidents and controversies in autonomous driving and various search recommendation services right now. Therefore, it is difficult to prove with the conservative level required by the nuclear field that the applications of AI and automation do not bring as negative concerns as positive effects.

In order to technically respond to the challenge of the fundamental uncertainty of human factors and human errors after the TMI accident, Nuclear Power has achieved a nuclear system that can show that safety has improved to a new level by introducing the concept of the Man-Machine Interface System (MMIS) that manages the safety of the entire system including humans from the conceptual design stage. However, the Chernobyl and Fukushima accidents again require a new level of safety. This is because organizational and human factors, which have emerged as a safety culture of various stakeholders and organizations beyond workers, still acted as uncertainties. Now, in order for the next generation of nuclear power such as SMR to be successful by introducing AI and automation, a new safety paradigm and its technological methodology that can be fully recognized for safety at the forefront of the realization of the technology are required.

3. Five Limitations and Problems within the Existing Safety Paradigm

Traditionally, safety has been defined as "Freedom From Hazards." (IEC/ISO, ASME standards and guidelines, etc.) For the technical implementation of safety in practice, the concept of risk, a contrasting measure of safety, was actively utilized to replace minimization and zeroing of risk with safety practice tasks. However, there were fundamental limitations and problems in the practice of safety, which are experienced in various ways in the field of nuclear power. For the successful realization of the next generation of nuclear power, technology with sufficient persuasive power can be achieved by considering the experiences and new demands. The following is the latest issue of nuclear safety, and five starting points for the integrated safety paradigm necessary for the next generation of safety are summarized through discussions on the results of previous studies such as Survey and interviews with stakeholders on human errors and safety culture, and review of safety approaches in multiple units, etc.

3.1 Safety can be defined in a variety of Perspectives and Measures

Most technologies have specific measures to realize this according to an obvious perspective. Functional safety, which has rapidly developed from nuclear power and contributed greatly to safety, is a very important point of view, and there are well-developed measures (for example, *cdf* in the PSA field), but these alone should not realize the safety of the next generation of nuclear power. Reductionist views on AI and automation functions can be said to be inadequate beyond insufficient. The limitations of presenting software safety as the probabilistic overall reliability of defects in coding, or functionally approaching cyber or physical security risks are clear. Traditionally, nuclear power has suggested the probability of an extremely low frequency accident as an important basis for safety, but the event and its probability assumed unilaterally in the field of nuclear power in the first place mean only a very specific point of safety.

Therefore, it is necessary to develop a variety of safety measures that can represent the perspectives of all stakeholders involved in nuclear power. It is painful but important lessons that nuclear energy has encountered failure to communicate with stakeholders and rejection, as safety measures related to fatality and risk of cancer/genetic defects have not addressed areas such as radioactivity-related mortality and cancer/ genetic defects.

3.2 Safety is not static, but dynamic in itself

Everyone would agree that safety is dynamic and not static. However, technically, there are not enough ways to capture dynamic safety characteristics. Efforts to capture, evaluate, and improve dynamic safety from an individual technical point of view in the nuclear field

have been consistent. However, self-defined safety paradigms and standards have been developed and applied in individual applicable (deterministic safety) technologies. Although an integrated safety paradigm that encompasses the entire nuclear system is being implemented in PSA and others, it has not deviated from a simple extension of the very narrow or static safety paradigm depending on the issues at hand. For example, it can be seen in the example of trying to capture the reliability of a nuclear system at the operating stage based on a static PSA centered on system reliability, or focusing the safety of multiple exhales on expanding the safety of individual exhales.

Nuclear systems deal with safety as a prerequisite, considering the dynamic behavior of the system from the design stage, but the operational stage in which the actual behavior of the system takes place involves various factors, including various uncertainties that are not addressed by probability or determinism. The representative example was the human factors of workers recognized in TMI accidents, but later, in Chernobyl and Fukushima accidents, there was a perception that the safety culture of the organization and society was flawed.

Dynamic safety is a new paradigm that does not imply the safety of the system's dynamic state, but rather that safety itself has dynamic characteristics. Therefore, design, operation management, and licensing should be carried out on the premise of a continuously changing reality of safety throughout the life cycle. It is necessary to meet the requirements of the life cycle risk management system required by NIST in the United States and the AI Basic Act that recently took effect in Korea, but it will be necessary to manage it from the perspective of maximizing safety rather than from the perspective of minimizing risk.

3.3 Safety is Cognitive and Social beyond Objectives

In technology that has dealt with safety in reality, objectivity is a major premise, so objectivity is important in the actual content of safety. Safety-related data should be an objective and universal process and outcome that are sufficiently technical as the basis for dealing with safety. For example, the risk used as a practical measure of safety is defined as the product of a specific loss and its probability of occurrence. (Risk = Loss x Prob) Various data used here are derived through a rigorous validity estimation process from actual measurement (operation performance or experimental measurement, etc.). Therefore, safety is recognized as objective and absolute, so it is scientifically accepted as the only fact. The enormous effort of technology to deal with objective safety through the engineering safety paradigm is something that anyone can respect.

However, this is not the case in reality. Safety estimated by probability and the like is only an ideological one, but reality is that it is realized in a specific way to a specific object at a specific time. This is because objective estimation contains fundamental limitations of science that extrapolate and estimate the future based on past experiences. Even mechanical factors cannot confirm future conditions and behaviors as an extension of the past (and its trends), much less the combination of environment and nature, as well as related factors such as humans and organizations that determine the final realization of safety.

This is because the probability of rain is actually meaningless to the person who gets rained on. The actual meaning of probability and loss is what the person and society actually experience related to the event, so it must be converted into cognitive or social values.

3.4 Safety can be Positive beyond the traditional risk minimization

Traditionally, safety aims at minimizing and zeroing (loss) risk in practice. In nuclear power, minimizing the radiation leakage caused by core accidents and accidents is an important design and operation goal. The Rasmussen report, famous for dealing with full-fledged comprehensive safety in nuclear power, set the goal of lowering the mortality rate of life due to nuclear power below the doomed probability. Lowering the mortality rate has traditionally been a top priority paradigm in dealing with safety and has become the core of safety technology. Minimizing risk (and loss) is because the safety goal at hand is clear.

However, safety is not a negative value centered around risk which is fundamentally something to be minimized, but a positive value to be maximized such as *Resilience* and Safety II. Additionally according to the new paradigm of *Safety III*, it is possible to add a positive value of safety beyond minimizing the occurrence of immediate failure, defects and losses.

In nuclear power, all deviations to achieve this based on self-set goals have been treated as anxiety (i.e., risk). However, the safety achieved by nuclear power on its own is not limited to minimizing risks, but includes defining new values and realizing them. Just as the safety of ships at ports versus ships sailing far away in the ocean is fundamentally and quantitatively different, it will be possible to continuously develop active and proactive safety of nuclear systems to seek additional positive values of safety. This means that in the case of AI and automation, the possibility of adding various positive safety values can be explored in open-ended manner. In particular, when regulations and permits are made through safety evaluation, the additional safety of nuclear systems should be differently recognized as an extension of new safety values rather than being criticized on the premise of traditional completeness. (2025 Lee)

3.5 Safety is Ecological and constantly changing

Among the safety paradigms, the longest-lasting limitation is the limited perspective on safety. Confidence in safety can be the most serious problem that threatens safety in reality. Even if it is confirmed as social safety as well as engineering safety, safety should not be determined. As you can see from the saying, "Safety is Alive!" Safety is fundamentally ecological. Although you can be sure of being safe at a certain point and point of view, safety should not be sure of being definitive, but it should receive constant attention and attention.

In particular, it is very important to maintain uncertainty about safety in the case of non-regressive and catastrophic characteristics such as nuclear safety. Therefore, nuclear safety should be continuously monitored and managed in any case and better safety should be pursued. Safety is an important criterion for making specific engineering decisions in next-generation nuclear systems, but this is a minimum requirement and not a sufficient condition for safety, which the nuclear field emphasizes and must be continuously developed.

4. Approaches for New Safety Paradigms

In order to actively introduce AI and automation, a specific methodology that can overcome the limitations of the traditional safety paradigm is needed. The following is a proposal for a technical application of the proposed paradigm shift.

First, it is an extension of the safety measure. In the introduction of AI and automation, the reliability of failure of engineering functions and the safety measure of radiation exposure must be deviated. For example, in order to be recognized for AI safety, LG proposed to verify 226 comprehensive risk possibilities, such as Augmented Universal Taxonomy (UT). Therefore, it is urgent to develop a taxonomy covering the risk possibilities expected by the introduction of AI automation systems in the nuclear field. However, the traditional safety of this risk perspective should be constantly supplemented and expanded in that it is a necessary and not a sufficient condition

Second, securing dynamic safety through AI and automation. AI and automation can also be applied to the functional advancement of nuclear systems, but they can be used to make up for the remaining margins of safety or to dynamically manage, respond to, and supplement safety. Everyone knows that safety is dynamic, but technically there are not enough ways to capture and take responsibility for dynamic safety characteristics. The basic paradigm for this should be based on the perspective of a joint-collaborative system that takes into account cooperation with humans in the introduction of AI and automation. It must meet the requirements of the life cycle risk management system required by the AI Basic Act, which took effect in Korea recently, but it is difficult to achieve the purpose by applying fixed safety standards defined in the (deterministic safety) field from the perspective of individual technologies. A new paradigm is needed that dynamically applies various detailed safety measures to dynamically change and track the safety of the entire system. In order to satisfy the continuously changing safety over the life cycle, a design, operation management, and licensing system should be developed that can deal with the substance to manage it through specific safety standards.

Third, a new quantification methodology for safety and risk must be applied. Traditionally, the engineering calculation method of risk has been used, equating safety with no risk. The simple arithmetic formula of risk called the product of probability and loss (Risk = Loss x Prob.) requires the addition of a process of converting into cognitive and social values, and behavioral economic methodology can be used immediately. In the safety evaluation of AI and automation, a more realistic fanfare scale can be applied using behavioral science correction formulas and integrated calculation formulas beyond engineering calculations. A recent proposed approach from the perspective of behavioral science (2022/2023 Lee) to nuclear safety is the fundamental starting point.

$$\text{Perceived Risk } (R') = f\left(\{ u(\text{Loss})_i \times \pi(\text{Pro.})_j \}_k \right)$$

- ✓$u(\text{Loss})_i$ = utility value of Loss_i
- ✓$\pi(\text{Prob.})_j$ = weighted prob. of Pro_j
- ✓$f(\text{Risk}_k)$ = integration of Risk_k

'u' means utility function that might be convex for gain and concave for loss along the reference point selected by people in risk perceptions and decisions. 'ð' means decision weight that may be a typical s-shape curve of conservatism∫ means the integral of risks rather than simple additive calculation.

Fourth, it is necessary to maximize safety, not minimize risk. There are already numerous safety functions designed to reflect the perspective of resilience, but their meaning has been treated with a narrow perspective of preventing risks. AI and automation are technologies that can show very excellent advantages in resilience such as monitoring, maintenance, management, and recovery of safety, so they can be actively used to maximize safety. For this, it is essential to apply the Safety II perspective. Safety can be maximized not from failure and loss, but from the perspective of the sustainability of a safety state without failure and loss. It can also be applied to maximizing positive safety (value) such as Safety III. For example, AI and automation technologies will be able to show clear contributions to various types of positive safety, such as surplus safety, restoration safety, and additional safety. Followings are a few examples of positive safety items over traditional negative safety. (2024/2025 Lee)

⋆ *S3.1 Surplus/Static Safety*
⋆ *S3.2 Effortive/Dynamic Safety*
⋆ *S3.3 Additive/Challenge Safety*

Fifth, it is necessary to develop a system that continuously maintains safety. Achieving engineering integrity that completely eliminates the risks of AI and automation is only ideal and impossible. Basically, safety is considered incomplete, and it is necessary to develop new areas of monitoring and control functions that can include the culture of organizations and stakeholders. AI and automation will not only complement the functions of existing nuclear systems, but will also open the possibility of new functions.

5. Launching of a New Safety Paradigm with AI and Automation in Nuclear

In the next generation of nuclear power that introduces AI and automation, the methodology discussed and proposed above is a simple starting point for obtaining the practical effect of the new positive safety paradigm, and the practical difference will be subtle until the full introduction changes the existing safety approach and concrete results are obtained. The limitations trapped in the negative safety of this traditional safety paradigm and the problems that fundamentally cannot be *Zeroed* have not yet been properly recognized or overcome even in the relevant specialized fields such as safety engineering and safety. However, in the field of nuclear power, it has the potential to organize and overcome. This is because, for practical reasons, the nuclear power field, which has already met all the safety needs of various levels, already includes a positive safety paradigm, and if it is clearly organized, it has the potential to systematically satisfy the new requirements of social and technological development.

Nuclear safety cannot be achieved with traditional safety paradigms such as AI basic laws or risk classification system-based verification in designing and developing new systems that have adopted AI and automation in the realization of next-generation nuclear systems at hand, such as SMR or nuclear waste technology-related systems, such as design and development programs, safety evaluation and licensing, and securing public acceptance. It is expected that effective realization of AI and automation can be achieved by voluntarily and proactively presenting and realizing the new safety paradigm proposed through the five directions in this study. This will enable the nuclear sector to continue its history of developing and introducing new safety paradigms more actively than any other field, developing technical methodologies for specific practices, and actually building safe systems in the future.

*** Acknowledgement**

This paper was written by revising and supplementing the previous two presentation papers ("A Direction of Human Error Research According to the Extended Concept of Safety" (2024/2025 Ergonomic Society of Korean) and “A Positive Safety Paradigm Proposed for Advances in Human Factors Engineering and Human Error Studies in Nuclear” (KNS-2025 Autumn, 2025))

References

- Kahneman, D.& Tversky, A., Prospect Theory: An Analysis of Decision under Risk, Econometrica, 47(2), 1979
- Hollnagel, E., Synesis, : The Unification of Productivity, Quality, Safety and Reliability, Routledge, 2020
- Hollnagel, E., From Safety to Safely- Principles and Practice of Systemic Potentials Management, Routledge 2025
- IAEA, Considerations on Performing Integrated Risk Informed Decision Making, TECDOC-1909, 2020,
- OECD, Behavioral Insights and Public Policy: Lessons from Around the World, OECD, Paris, 2017
- Rasmussen, J., The Role of error in organizing behavior, Qual. Saf. Health Care;12, 377-383, 2003
- Thaler, R. , Mis-Behavior, 2023
- Thygerson, A.L., Safety : Concepts and instruction, Prentice-Hall, 1972
- Tversky, A. & Kahneman, D., Judgment under Uncertainty : Heuristics and Biases, Science 185, 1974
- Wickens, C.D., Engineering Psychology and Human Performance, 2nd ed. Harper Collins Pub., 1992

(Prior papers presented by Y.H. Lee)

- *Human Error 3.0* Concept for High-Reliability Era, *ESK-2015-Fall*, 2015
- A Proposal to Revise the Risk Concept and Approach based on Behavioral Science Perspective for Risk Communication and PA in Nuclear, *KNS-2018 Spring*, 2018
- A Behavioral Scientific Proposal to Revise the Multi-Unit PRA for Improving Risk Communication and Public Acceptance on Nuclear, *KNS-2020 Spring*, 2020
- A Brief on Nuclear Safety History and Various Safety Concepts for Social Acceptance of Nuclear Power: A Behavioral Science Perspective , *KNS-2024 Spring*, 2024
- A Preliminary Study on the Achieved and Extendable Concepts of Nuclear Safety to Improve the Social Acceptance in Korea, *KNS-2024 Fall*, 2024
- *Human Error 3.0*: How to Cope with Human Errors including Violations for Safe and Secure Future, *Proc. of International Ergonomics Associations IEA-2024*, 2024
- Challenges to Human Error Studies according to the Extended Concepts of Safety: for New and Clear Safety Future, *ESK 2024 Fall*, 2024
- A Critical Revisit to Human Error Studies for Advances in Human Factors Engineering: *Critiques to Traditional Safety Paradigms and Safety II, ESK 2025 Spring*, 2025
- A Discussion on the Development Direction of the Next Generation I&C System in Nuclear : To Cope with the Human Error Uncertainty Remaining after Fukushima Accident, KNS-2025 Spring, 2025
- A Positive Safety Paradigm Proposed for Advances in Human Factors Engineering and Human Error Studies in Nuclear, KNS-2025 Autumn, 2025