Envisioning Emergent Behaviors of Socio-Technical Systems Based on Functional Resonance Analysis Method

Hirose Takayuki

Envisioning Emergent Behaviors of Socio-Technical Systems Based on Functional Resonance Analysis Method

Takayuki Hirose
Supervisor: Professor Tetsuo Sawaragi
Department of Mechanical Engineering and Science
Kyoto University

A thesis submitted for the degree of
Doctor of Philosophy in Engineering

2020

Abstract

This thesis provides simulation models to envision emergent behavior, or more specifically, safety of artifacts as socio-technical systems, based on Functional Resonance Analysis Method (FRAM) and addresses potential problems inherent to three principles to design human-machine systems.

It has traditionally been believed that the safety can be ensured by identifying and eliminating unsafe factors. However, people came to realize that it is not enough since unsafe events such as hazards or accidents are still unavoidable no matter how carefully we brace for them. Specifically, those experiences suggest that there is a lack of balance between Verification and Validation (V&V) in their traditional approaches; the traditional safety management is very good at verifying how the artifacts should be designed, but essentially poor at validating how they can maintain their validity in real fields of practice.

One typical approach for this challenge is to carry out Work Domain Analysis (WDA) and investigate functional safety of target artifacts; the WDA enables us to model the targets and investigate their safety at functional abstraction level. This approach is expected to provide various insights which cannot be obtained at the physical level — the field of traditional safety management. The problem is that the investigation of the functional safety is generally qualitative, and it is therefore difficult to confirm its validity.

The objective of this research is to develop simulation models based on the FRAM and overcome this problem. In this respect, this thesis firstly summarizes the nature of difficulties in ensuring the safety of artifacts. Also, the historical context of the safety is reviewed to introduce the resilience engineering — one of the latest ideas of how to ensure the safety of artifacts. The focus is then shifted to the FRAM which has been proposed as a practical method of the resilience engineering; this thesis reviews an overview of the FRAM including its limitations and propose numerical simulation models based on the FRAM to make it more practical. In addition, this thesis addresses potential problems inherent to three principles to design human-machine systems, i.e., the Compensatory Principle, Leftover Principle, and Complementarity/Congruence Principle, through case studies with the proposed FRAM models.

The first case study examines the validity of the Compensatory Principle. In this principle, functions of humans and machines are separated and allocated to what they are good at, and it is pointed out that their roles often tend to be fixed even in unexpected situations. To examine the issue, this case study investigated an actual air crash accident with the proposed FRAM model. The FRAM simulation envisioned how the validity of the operation, especially focusing on the feasibility of operational procedures, had been changing during the accident sequence. The simulation result consequently confirmed that it is necessary for this principle to take into account the validity of such predefined function allocation in ever changing environment, and the FRAM model can be utilized for such kind of stress test.

The second case study addresses an inherent problem of the Leftover Principle. This principle is to automate everything found to be feasible by designers and push human beings to a domain where too complex tasks/activities to be automated are left. The operation of systems based on this principle shall generally depend on implicit knowledge or skills of human operators, and it is generally difficult to elucidate their validity. To address the issue, this case study examined the validity of an empirical knowledge currently inherited in the steel production industry; the simulation result provided several insights about why the knowledge can be effective and confirmed that its dynamics is closely related to that of complex systems.

The third case study is to demonstrate the importance of the Complementarity/Congruence Principle, or more specifically, human-machine collaboration for the future. It is expected that more and more automations will be introduced into our daily lives, and some of them are designed to reduce the involvement of human beings as much as possible. However, the history shows that such kind of automations often confuses human beings, and the situation could be even worse since such the consumers are generally "novice" of the automations, contrary to "professionals" such as aviation pilots. To demonstrate this problem, this case study examined the feasibility of the SAE conditional driving automation in time-critical situations. The result suggested that human drivers must be involved in driving activities even if the highly automated driving systems are responsible for major part of the driving tasks.

In conclusion, this thesis points out that it is difficult for traditional reductionism or methodologies of reliability analysis to provide these kind of insights; the functional safety must be investigated for this reason, and certain methodologies to support the investigations are required. The proposed FRAM models are one possible solution.

Acknowledgement

First and foremost, I wish to express my deepest gratitude to my supervisor, Professor Tetsuo Sawaragi, who continuously supported me on every activities as a researcher. He always inspired me with his profound knowledge and encouraged me throughout this research. He also provided great opportunities to see and discuss many people across the fields all over the world. These experiences shall greatly support me from now and forever.

I am also thankful to all staffs and members in Sawaragi laboratory. Especially, Lecturer Hiroaki Nakanishi and Assistant Professor Yukio Horiguchi always provided keen insights and comments to convince me of various perspectives of this work. Besides, Secretary Ms. Minato supported me by taking care of all the paperwork in the university and encouraged through daily conversation. The experience in this laboratory would not have been great without them.

This research is supported by many people outside of Kyoto University as well. Dr. Hiroshi Narazaki, an engineer of Kobe Steel, Ltd., shared a lot of insightful knowledge and suggestions based on his profession. The members of Total Flight Operation System Study Group (TFOS.SG) shared a number of valuable information which cannot be obtained in the academic field alone; their airmanship inspired and motivated me a lot. I also had very constructive discussions with the members of FRAMily. Especially, Professor Erik Hollnagel at University of Southern Denmark, Honorary Professor David Slater at Cardiff University, and Researcher Riccardo Patriarca at Sapienza University of Rome, took time for me and shared a lot of ideas about future prospects of FRAM. Moreover, Dr. Hideki Nomoto and Researcher Yasutaka Michiura, engineers of Japan Manned Space Systems Corporation (JAMSS), provided me opportunities to have regular discussions, and furthermore, a post in this company as a researcher after my graduation. It is indeed difficult to mention about all the people here, but I am so gratefull to everyone who got to know with, inspired, and supported me through this research.

In the end, I take this opportunity to express my profound gratitude to my family. They alaways respected my ideas and encouraged me to try Ph.D. It had been impossible for me to complete this work without their dedicated support and patience.

Contents

Chapter 1

Introduction

1.1 Envisioned World Problem: Difficulties to Predict Impacts of New Technologies

Innovative technologies generally involve a trade-off between convenience and safety. They have been providing more power, speed, or energy with human beings, for instance, and contributing to the development of our society. On the other hand, their evolution is currently so rapid and significant that they often go beyond our capacity to control them, resulting in unexpected consequences. The dual nature of the technologies, or more generally, artifacts is always the case when we design, develop, and use them.

Automation, for example, is originally introduced to reduce the workload of human operators and improve their accuracy of tasks/activities. It has already been implemented in many fields such as transportation, nuclear power plant, or manufacturing industries and surely contributing to their safe and efficient operations. However, negative aspects about the automation have been pointed out for decades as well. Carroll and Campbell (1989), for example, pointed out that tasks/activities carried out by human operators are changed by introducing the automation, and it also has some effect on others. In other words, the operational tasks/activities are so interdependent that it is impossible to completely isolate a certain part of them and replace it with the automation, which is also known as the substitution myth (Sarter et al., 1997). Also, Endsley (1995a,b, 2001), Sarter and Woods (1995), and Degani (2004) unanimously suggested that the automation demands human operators carry out more complex cognitive tasks/activities; the demand could cause the loss of situation awareness of the operators and lead to discrepancies between the situation awareness of the operators and actual behaviors of the automation. The operators

are consequently confused with unexpected behaviors of the automation, known as automation surprise (Sarter et al., 1997), and the automation surprise could result in human errors in the worst case. Those negative aspects suggest that the automation without elaborate consideration of the design could ironically degrade the safety of the operations contrary to its original purpose. In the current context of technological development, almost all innovative technologies can be associated with the automation, and the safety issues about the automation are thus becoming more and more critical in the society with increasing its speed.

Safety of the artifacts is, needless to say, carefully investigated in their Research and Development (R&D) processes, and safety measures are conducted to eliminate risks identified in the processes as much as possible in advance. However, the history of the safety suggests that there are limitations about the traditional approaches. Specifically, the real world is so complex that it is almost impossible to describe the whole aspects in early phases of the R&D processes. This problem has actually been causing major symbolic events such as Controlled Flight Into Terrain (CFIT) of American Airlines Flight 965 (CAPA, 1996; Simmon, 1998), landing accident of China Airlines Flight 140 in 1994 (JTSB, 1996), or successive two crashes of Boeing 737 MAX operated as Lion Air Flight 610 and Ethiopian Airlines Flight 302 in 2018 and 2019, respectively (KNKT, 2019; AIB, 2019). The detail of these cases will be presented in the succeeding chapters, suggesting a gap between how the artifacts, even including Standard Operational Procedures (SOPs) or training programs, are designed to function in the R&D processes and how they actually function in real fields of practice. In other words, the traditional safety management is very good at verifying how the artifacts should be designed, but essentially poor at validating how they can maintain their validity in real fields of practice. This is basically not due to the fault or carelessness of human beings but the intrinsic difficulties to carry out the proper Verification and Validation (V&V) processes.

Woods and Dekker (2000) point out the nature of this problem as shown in Fig. 1.1. According to them, the introduction of a new technology or system is not a simple substitution of machines for people but an intervention into ongoing fields of activities. Specifically, the introduction of the new technology is not a simple manipulation of a single variable but a change that reverberates throughout a system transforming judgements, roles, relationships, and weightings on different goals. The new technology in this context can be regarded as not only an object, but also a hypothesis about how technological change will transform practice and shape how

people adapt, and designers of the new technology function as experimenters to generate the hypothesis in this sense. In addition, the hypothesis is usually validated through actual operations by practitioners in a real field of practice. It is therefore essential to anticipate the impact of the new technology involving human factors, and it should, in theory, be addressed in the early phase of the R&D process. However, the human factors are often pushed to the tail of the R&D process. This is because it is quite challenging to envision the lucid visions of the evolving field of practice emerging out of the "non-linear" interactions among human, machines, and their surrounding environment; practitioners adapt to difficulties, re-shaping artifacts to function as tools to meet the demands of the field of activity with their creativity. Besides, empirical tests in this phase generally provide too little information at too high a cost for further improvements. The difficulty to anticipate full-range of potential effects or post-conditions of technological change is known as the envisioned world problem (Woods and Dekker, 2000; Woods and Christoffersen, 2002), which has been causing the unbalance of the V&V processes. One of the most considerable demands to conduct the proper V&V processes is, therefore, to develop innovative methodologies and overcome the envisioned world problem.

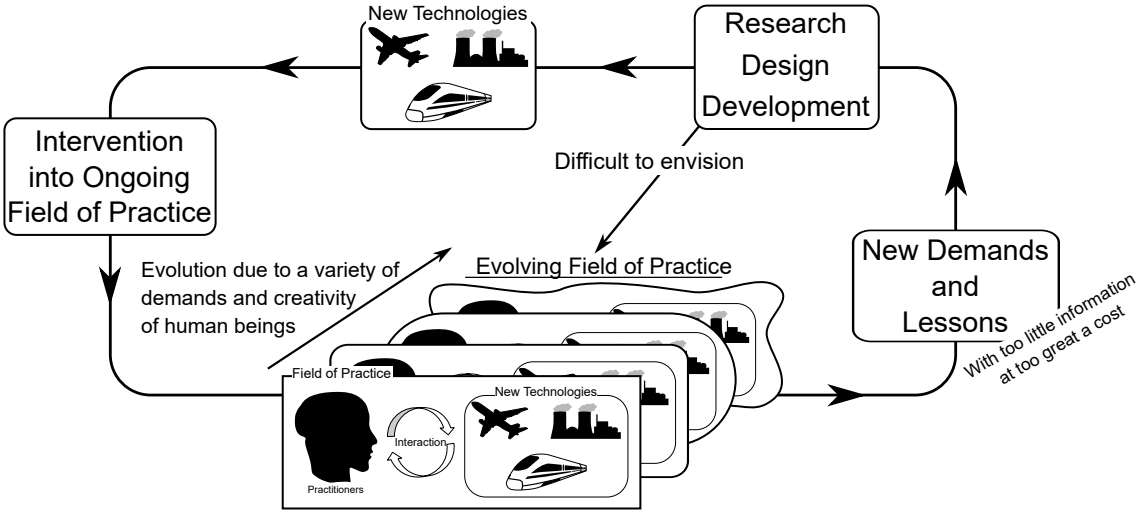


Fig. 1.1: Envisioned world problem: difficulties to anticipate impacts of new technologies.

1.2 Approach for Envisioned World Problem

The nature of the envisioned world problem is how to build a model of the evolving fields of practice and investigate its feasibility in specific situations. In other words,

the focus of the problem is put on how cognitive actions of human beings are carried out under a variety of situations and how they result in a specific consequence. However, the real world is so complex that it is intrinsically difficult for real-oriented or physical approaches alone to address all aspects of the issue. Besides, insights obtained from such approaches are often too ad-hoc to be practical, which is also consistent with the limitation of empirical tests in R&D processes. The envisioned world problem therefore requires additional approaches to build and analyze models of the fields of practice from the perspective of complex human-machine systems.

Work Domain Analysis (WDA) which is a dimension of Cognitive Tasks Analysis (CTA) have conventionally been conducted in this context. According to Roth and Mumaw (1995), methodologies of the CTA have been developed to identify and take into account the cognitive processing requirements inherent in performing complex tasks. In other words, they identify the dimensions of task complexity and the demands imposed by the surrounding environment of the field of practice.

The typical approach of the WDA is to model target systems at multiple abstraction levels so that those models characterize the purposes for which the engineered system has been designed, and the means structurally available for achieving those objectives. These models enable us to investigate the structural constraints of the goal-means relationships inherent in a variety of situations. A variety of methods, such as the Abstraction Hierarchy (AH) by Rasmussen (1986) or Multilevel Flow Modeling (MFM) by Lind (2011), have been developed to map out the range and complexity of tasks inherent in a field of practice. They are expected to provide the basis for defining control requirements, information to display, and roles that humans will play in the human-machine interactions.

Based on the idea of the WDA, one possible solution for the envisioned world problem can, therefore, be summarized as shown in Fig. 1.2. The nature of the solution is to leave the real-oriented physical world once and approach the problem from a higher — functional abstraction level. Specifically, the abstraction process can be conducted with certain WDA methodologies, and the safety of the functional models, i.e., functional safety of target systems is analyzed. The functional safety analysis consequently provides a variety of generic insights with us, which cannot be obtained with physical approaches alone. In the end, the ultimate goal to overcome the envisioned world problem is to create a structure in which both the physical and functional processes shown in Fig. 1.2 are actively cycled so that the R&D processes can get more feedbacks from both of the cycles; this research aims at constructing the functional part of the structure and investigating the functional safety.

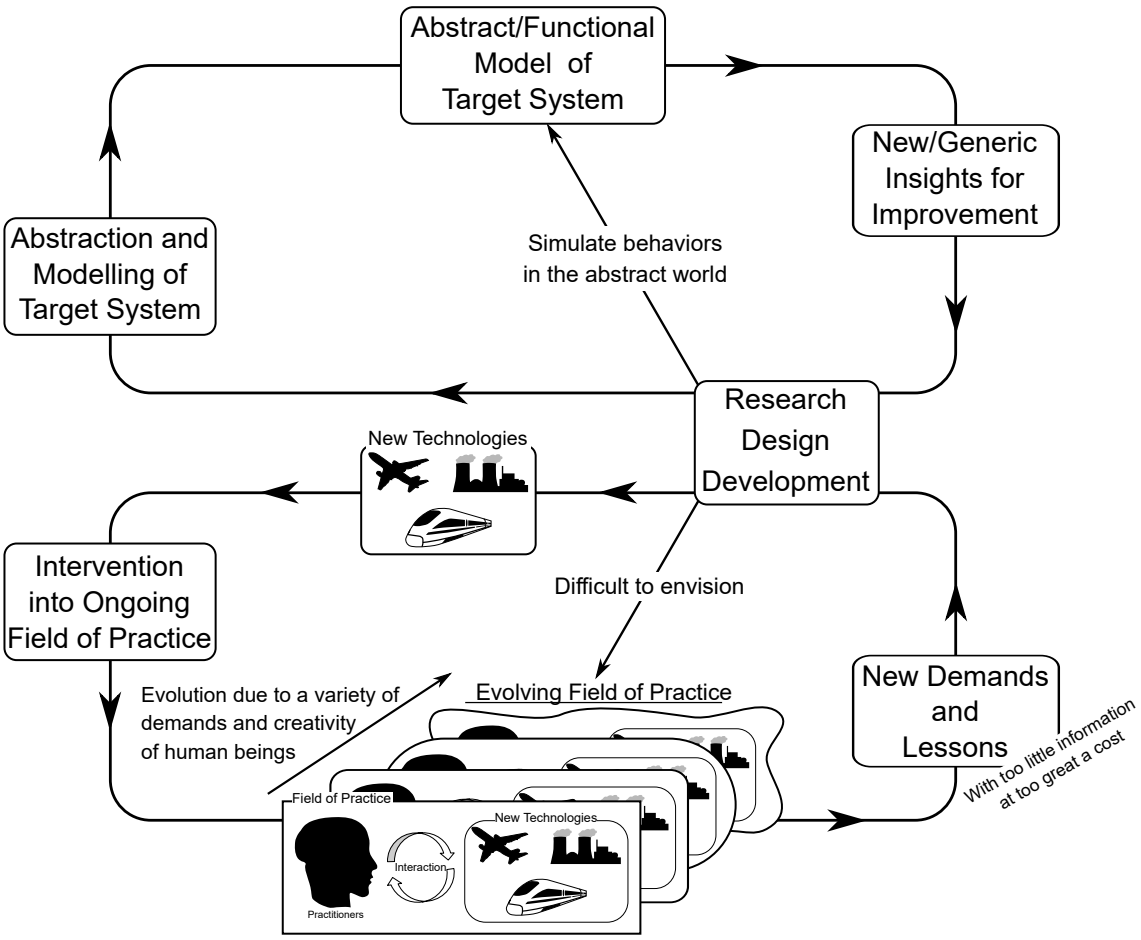


Fig. 1.2: Approach for envisioned world problem: functional safety analysis of human-machine systems.

1.3 Overview of This Research

The goal of this research is to propose a framework of the functional safety analysis for proper V&V process, and moreover, future safety of artifacts. The problem is that traditional methodologies of the functional modeling/analysis based on the WDA have some limitations. According to Roth and Mumaw (1995), the traditional methods are to structure analyses of operator information, control requirements around specific event sequences, and preplanned response strategies for handling those event sequences. The traditional approaches are, therefore, useful in ensuring the Human-Machine Interfaces (HMI) support of human operators in pre-analyzed situations. However, they do not provide a principled way to identify information and control requirements to support operator performance in unanticipated situations. In addition, the functional safety analysis is generally qualitative, and it is therefore difficult

to confirm its validity. One of the primary objectives to achieve the research goal is, therefore, to develop methodologies and overcome these limitations.

In this context, Functional Resonance Analysis Method (FRAM) has been proposed by Hollnagel (2004, 2012b) as a practical methodology of the resilience engineering (Hollnagel et al., 2006; Hollnagel, 2017). The resilience engineering and FRAM have been developed to investigate the complex safety of socio-technical systems which is an extended framework of the human-machine systems. They are now attracting their attentions since their focus is put on 1) why things go well rather than why things go wrong, contrary to the traditional concept of the safety, 2) the complex behavior of the socio-technical systems emerging out of the "non-linear" interactions among the various components making up the systems, and 3) their ability to maintain acceptable performance under expected and unexpected conditions alike. However, they are currently so conceptual that many people in both academic and industrial fields are eagerly working on how to make them practical as well.

The objective of this research is, therefore, to develop a practical simulation model based on the FRAM and addresses potential problems inherent to three principles to design human-machine systems. The remainder of this thesis consists of the following chapters. First of all, the chapter 2 and 3 provides the background and theory related to the resilience engineering and FRAM: the chapter 2 reviews the historical context of the safety and introduce the concept of resilience engineering; the chapter 3 summarizes the overview of FRAM and proposes its numerical simulation models. Then, the chapter 4 – 6 addresses potential problems inherent to three principles to design human-machine systems, i.e., the Compensatory Principle, Leftover Principle, and Complementarity/Congruence Principle (Hollnagel and Bye, 2000; Hollnagel and Woods, 2006), through case studies with the proposed FRAM models: the chapter 4 provides a validation of the Compensatory Principle, focusing on the dynamic feasibility of a flight-deck procedure carried out in the sequence of an actual air crash accident; the chapter 5 shows an investigation of a problem inherent to the Leftover Principle through an analysis of empirical knowledge currently inherited in the operation of steel production lines; the chapter 6 discusses the importance of the Complementarity/Congruence Principle through validation of the SAE conditional driving automation in time-critical situations. In the end, the chapter 7 reviews the impact of the functional safety analysis or V&V processes based on the FRAM simulations and concludes this paper with proposals for the proper design of artifacts and co-creative human-machine interactions to effectively cope with new technologies.

Chapter 2

Resilience Engineering in Historical Context of Safety Management

2.1 Concept of Safety

One of the best-known definitions of safety is the following, which is often referred to across other industries and professions:

> Safety. The state in which the possibility of harm to persons or of property damage is reduced to, and maintained at or below, an acceptable level through a continuing process of hazard identification and safety risk management (ICAO, 2009).

The traditional concept of the safety is therefore associated with a state that is free from harm without the consequences of failure, damage, accidents, or other undesirable events. Also, the purpose of traditional safety management is to identify and eliminate things that go wrong, i.e., hazards, or adverse events as much as possible to prevent unwanted outcomes. However, this sounds somewhat strange from the perspective of science because the "existence" of safety is paradoxically proved by the "absence" of unacceptable consequences, according to Hollnagel (2017). The concept of safety has been changing, as if in response to this paradox, as we experienced the technological developments, complications of the society, and lessons learnt from their sacrifices. The resilience engineering has been developed in this context.

This chapter, first of all, reviews the historical context of the safety concept from three perspectives, i.e., development of contributing factors to the safety, accident models, and Human Reliability Analysis (HRA). The review is then extended to the introduction of the resilience engineering.

2.2 Contributing Factors to Safety Suggested by Historical Major Accidents

2.2.1 Safety Supported by Technical Factors

Accidents occurred until the middle of the 20th century were generally caused by technical factors such as malfunction of machines or destruction of structures. Specifically, the collapse of Tacoma Narrows Bridge, mid-air explosion of Comet, brittle fracture of Liberty Ships are known as typical examples of such accident. They are frequently highlighted because all of them include important lessons related to engineering aspect as shown below:

Collapse of Tacoma Narrows Bridge (Nakao, n.d.)

Tacoma Narrows Bridge was constructed at a strait of Tacoma city, Washington in the United States in 1940. The latest technology at that time was involved in its construction. However, the bridge collapsed only 4 months after its opening. The collapse was caused by wind-induced oscillations, whose wind velocity was just $19m/s$ at that time. This accident is currently known as a trigger of the fluid oscillation field to develop.

Mid-air Explosion of Comet (Kobayashi and Terada, n.d.)

Comet is the first-jet airliner with a pressurized cabin in the world, designed by the de Havilland Aircraft Company Ltd. The first flight of Comet was in 1949, and the Comet had been getting good reputations for its speed and comfortableness. However, three Comets crashed only within 2 years after starting its service, whose cause was unknown at that time; one major reason for this is that the method for investigation of air-crash accident had not been established yet. After a large-scale and exhaustive investigation, the direct cause was found to be some fatigue cracks of the fuselage caused by repetitive pressurization cycles. The investigation elucidated the mechanism of metal fatigue, which is well and widely known today, systematically and also contributed to the development of investigation methods of aircrash accidents.

Brittle Fracture of Liberty Ships (Kobayashi and Onoue, n.d.)

At the middle of the World War II, the United States constructed supply ships (DWT 11000 Liberty Ships) with a certain method to meet the demands of the Pacific War. The method was to join the parts of ship hull by welding instead of riveting, and the

construction period at docks had been greatly improved from six months to just four days. However, more than the 200 out of 2700 Liberty Ships sank or seriously damaged; some of whose body were cut and separated in no time as if glasses broke. This accident was directly caused by occurrence and development of brittle crack, which is brought about by the immature steel making technology and welding technique. This accident can consequently be regarded as the most expensive and huge scale experiments of the century (Kobayashi and Onoue, n.d.), having led to the development of steel making and welding technology.

Major accidents in the first half of the 20th century were occurring in this way. At that time, technical knowledge which is currently obvious was lacking. In addition, the structure of artifacts were simpler than that of today. That is why the main approach to ensure the safety was to avoid malfunction of machines or destruction of structures. However, the situation has changed around the middle of 20th century because of the dramatical development of computers and automation technologies.

2.2.2 From Technical Factors to Human Factors: Impact of Automation and Growing Importance of Effective Human-Machine Interactions

In the 1950s to the 1960s, the performance of computers had greatly been improved, and automation technology based on the computers were introduced to real operations of systems such as aircrafts or nuclear power plants. The eleven Levels of Automation (LoA) are currently defined by Sheridan (1992) and Inagaki et al. (1998) as shown in Table 2.1, and the latest automated systems can perform their tasks with the high levels such as LoA 8 or 9. However, the automation intrinsically involves the negative aspects briefly described in the chapter 1. One of the breaking triggers to expose the complexity and clumsiness of the automation is the nuclear disaster of Three Mile Island (TMI) occurred in 1979 (Cantelon and Williams, 1980); lessons learnt from the disaster convinced people of the importance of effective human-machine interactions.

Table 2.1: Each Level of Automation (LoA) and their definitions (Sheridan, 1992; Inagaki et al., 1998).

LoA	Definition
1	The computer offers no assistance; human must do it all.
2	The computer offers a complete set of action alternatives, and
3	Narrows the selection down to a few, or
4	Suggests one, and
5	Executes that suggestion if the human approves, or
6.5	Executes automatically after telling human what it will do; no veto is allowed, or
6	Allows the human a restricted time to veto before automatic execution, or
7	Executes automatically, then necessarily informs humans, or
8	Informs him after execution only if he asks, or
9	Informs him after execution if it, the computer, decides to.
10	The computer decides everything and acts autonomously, ignoring the human.

Nuclear Disaster of Three Mile Island (Cantelon and Williams, 1980; Chiles, 2002)

The TMI disaster was initially triggered by the break of a water supply pipe running through a plant, i.e., TMI-2. Just before the break of the pipe, an operation was carried out in order to relieve the clogging of resin in a condensate desalination tank, and very little water accidentally reached a control system of a valve to control all condensate demineralizer of TMI-2. The automated monitoring system judged this water as abnormal and abruptly shut down all water supply valves. The impact of this shut down broke a pipe in the turbine building, and the water supply was stopped. The core was exposed to the air as a result of this and melted down in the end. At this time, the warning system was launching more than 100 alarms, and the operators were confused with huge amount of information coming from the control panels.

It is of course possible to describe this accident from the perspective of technical factors. However, what is more important here is that the confusion of operators made the situation worse even though the warning system itself was "correctly" working as designed. This accident suggests that the safety can no longer be ensured without correct human-machine interactions, and it is insufficient to just improve the reliability of human or machines independently.

We have been facing the safety issues related to the effective human-machine interactions since then. Especially, a variety of cases can be seen in the aviation field that has a long history of the automation; some parts of their operations have already been automated for commercial use at least no later than 1970s, and it is now

even possible to complete the entire flight sequence from take-off to landing without operating control equipment such as yokes or thrust levers. On the other hand, the field has been experiencing a variety of accidents related to human factors and the automation. Specifically, the impact of Controlled Flight Into Terrain (CFITs) — accidents which does not involve malfunctions of aircrafts has been significant, suggesting that correct responses of human operators to mechanical systems play a significant role in the safety even if there are no physical/technical problems with the machines as well as the TMI disaster.

For example, one of the best-known CFITs is a landing accident of China Airlines Flight 140 at Nagoya airport in 1994 (JTSB, 1996), whose profile is shown in Fig. 2.1. The accident sequence started with the erroneous engagement of Go Around (GA) mode by Pilot Flying (PF) during the final approach to the airport. This engagement triggered the autopilot to climb, but the PF forcefully continued to descend by pushing the control column without disengaging the GA mode. This caused the confliction between the PF and autopilot as if it were a tug of war, which physically appeared at the elevator and horizontal stabilizer as shown in Fig. 2.1. The PF eventually gave up descending and increased the thrust of the engines to climb again. However, this resulted in abnormal pitch attitude and steep climb of the aircraft as if the one side of the tug of war, i.e., the PF suddenly released the "tension" of confliction. The flight 140 stalled and crashed into the ground in the end.

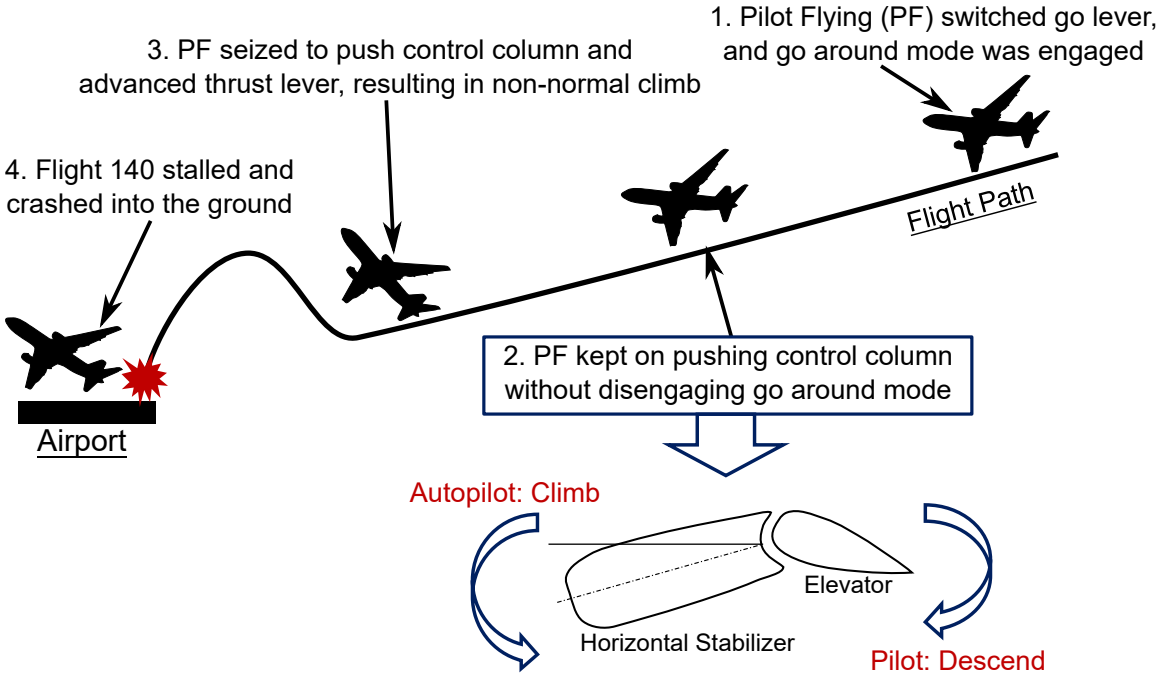


Fig. 2.1: Profile of accident: crash of China Airlines Flight 140 at Nagoya in 1994.

The importance of human-machine interactions, or specifically, human factors has been investigated as a result of lessons/demands learnt from these cases. The negative aspects of the automation which is briefly described in the chapter 1 are ones of the most famous and critical issues for them, and the problems are still alive even 50 years after the disaster of TMI-2. The problem is that the human operators are left in the domain where demands human operators perform more complex cognitive tasks/activities than ever as the automation expands its capability, causing their loss of situation/mode awareness and the automation surprise (Sarter et al., 1997); Endsley (1995a,b, 2001), Sarter and Woods (1995), and Degani (2004) have unanimously addressed the problem from their perspectives in this context.

Endsley (1995a,b, 2001) discussed the problem from the perspective of Situation Awareness (SA). She points out that human operators usually remain in automated systems as monitors to check if the systems perform properly and to detect non-normal conditions, nevertheless prevailing approaches to design systems is to reduce the workload of human operators (Endsley, 1995b). In addition, the automated systems generally provide huge amount of data with human operators, demanding the human operators process and utilize the too much data in the operation of automated systems. However, human operators are basically not suitable for these kind of tasks/activities; the data coming from the systems often makes the operators overloaded, which could lead to human errors in the worst case. In other words, the human being is not the cause of these errors, but the final dumping ground for the inherent problems and difficulties in the technologies we have created (Endsley, 2001). She put a focus on the importance of the SA in this context and proposed its model (Endsley, 1995a); the SA model consists of the three SA levels shown in Table 2.2, and the establishment of each SA level leads to the higher SA levels to support Decision Making (DM). One of the major discussion points of this model is how to identify which levels to put a high priority depending on a specific situation; it might be effective to present information directly supporting the higher levels of SA, i.e., Lv. 2 or Lv. 3 in time-critical situation, for example.

Table 2.2: Levels of SA.

Level	Description
1. Perception	To detect some data or elements of what is going on
2. Comprehension	To integrate the elements and comprehend the current situation or the reason why it is happening
3. Projection	To anticipate future status based on the comprehension of the current situation, leading to DM

Sarter and Woods (1995) addressed the problem from the perspective of modes executed by automated systems. According to them, it used to be easy to track the mode status of the automated systems before since their behavior was dependent on operators' input, and their response was quick and clear as well. However, the situation has changed by the development of new technologies; the number of modes has been increasing, and they provide human operators with a large number of functions and options to carry out a given task. In addition, the modes can be engaged by multiple ways and even activate themselves based on the surrounding environment states, and moreover, the modes can carry out the long sequences of tasks autonomously without noticing the human operators once they are engaged. This demanded operators know more about the systems than before to meet the new monitoring and attentional demands to track the mode currently engaged and its behavior. It is therefore difficult to maintain the awareness of human operators about modes of more automated systems, and it consequently leads to the loss of mode awareness, the inadvertent activation of an incorrect mode, and automation surprise.

Degani (2004) also discusses about the problem of the automation based on gaps between the awareness of mechanical systems' state by human operators and their actual states. The current systems, including the automation, are so complicated that it is almost impossible to fully understand what is going on in the systems. Proper design of HMI therefore plays a significant role in the comprehension of the complicated systems' state by the human operators; the role of HMI is to integrate and abstract data processed in the systems to display operational information so that the human operators can intuitively construct their appropriate mental models of the systems' state. However, the improper design of HMI could simplify, abstract, or conversely complicate actual states of the systems too much. This causes discrepancies of the systems' states between in understanding of the human operators and the real world, leading to the automation surprise as well as other issues.

In summary, the human factors or effective human-machine interactions are currently playing a significant role in the safety of artifacts. There are many challenges to design proper human-machine systems as shown in the above. In addition, the situation is getting more and more complex without waiting for the solutions. This is mainly because of the accelerating technological developments and complication of the society, especially after the 2000s. The growing complexity requires further expansions of the scope of the safety, which is beyond the framework of the human-machine systems.

2.2.3 Growing Complexities: Safety of Socio-Technical Systems

Features of recent accidents are somewhat different from what they used to be; although the human factors still play a significant role in the safety, it is no longer sufficient to describe current accidents within the framework of the human-machine systems alone. There are a number of additional factors involved in the safety. In addition, their complex interactions could result in catastrophic consequences, whose process cannot be explained based on cause-effect relationships or "linear" approaches. The following accidents are typical examples for such cases.

Amagasaki Derailment Accident (JTSB, 2007)

This is an fatal derailment accident occurred at a curve between Tsukaguchi and Amagasaki stations, Hyogo prefecture, Japan in 2005. This accident was directly triggered by the overspeed resulting from the driver's effort to recover a delay of the train schedule. However, the process having led to the overspeed was quite complicated. The driver was extremely afraid of the delay because such kind of "fault" must have had been reported to the company at that time, and there was a possibility for him to get a penalty, i.e., harsh and humiliating retraining programs known as *nikkin kyoiku*. JTSB (2007) found such organizational system to be a contributing factor to this accident; individual factors such as the overspeed of the train or "human error" of the driver cannot simply be blamed for the cause of this accident.

Fukushima Daiichi Nuclear Disaster (IAEA, 2015)

This is one of the most catastrophic nuclear disasters in history occurred in 2011. The trigger of this disaster was the huge earthquake and tsunami struck the east coast of Japan; the tsunami struck the power plant brought about flood, and backup generators were fatally damaged. The damage of the generators caused the loss of cooling systems and resulted in nuclear meltdowns accompanying hydrogen explosions in the end. However, the actual process leading to the catastrophic consequence was far more complex than described in literature. Many factors, including organizational or even political factors, were involved in the process, supposed to have interfered with the accident response processes.

Beside the catastrophic consequences, this accident also has another remarkable aspect. Yoshizawa et al. (2016), whose first author was working as a manager of the reactor unit 5 and 6 at that time, point out that the achievement of the cold

shutdown of those units was greatly supported by dedicated responses carried out by all the people working there. Moreover, he experienced that those responses were beyond prepared organizational rules by personnel to meet the situation with limited resources and uncertain information.

Landing Crash of Asiana Airlines Flight 214 **(NTSB, 2014)**

This is a fatal landing crash accident as a result of a failure of final approach at San Francisco international airport in 2013, whose profile is shown in Fig. 2.2. At the beginning of this sequence, Asiana Airlines Flight 214 was flying well above the desired glide path. The flight 214 did not initially correct the deviation, and altitude on the Mode Control Panel (MCP) was set to $3,000$ ft to prepare the autopilot in case of the go around is needed. After about one minute later, the Pilot Flying (PF) switched the Autopilot Flight Director System (AFDS) pitch mode to Flight Level Change Speed (FLCH SPD). This change of mode results in the flight 214 initiating a climb to the $3,000$ ft which was previously set. To prevent this unexpected climb, the pilot disengaged the autopilot and retarded the thrust levers to idle. This caused another change of mode: the auto throttle transitioned to the HOLD mode which does not control the airspeed. The flight 214 descended throughout the desired glide path after this, and the pitch attitude was steadily increased by the PF so that the aircraft maintain the desired glide path. However, the airspeed was no longer controlled at this time, and the flight 214 was flying at well below the glide path, whose airspeed was 15 knots below the desired approach speed. In the end, the Pilot Monitoring (PM) advanced the thrust lever and declared go around just before the runway of the airport, which was too late, and the underside of the aircraft struck the seawall of the airport.

This accident is also characteristic in the sense that multiple factors, not just limited to the human factors, are involved in during the sequence of the crash, which is also pointed by the official accident report (NTSB, 2014). The accident report concluded that the probable cause was the flight crew's mismanagement of the airplane's descent during the visual approach. Also, it blamed other factors such as the complexities of the autopilot system, the inadequate collaboration among flight crews, and inadequate training of crews for contributing factors.

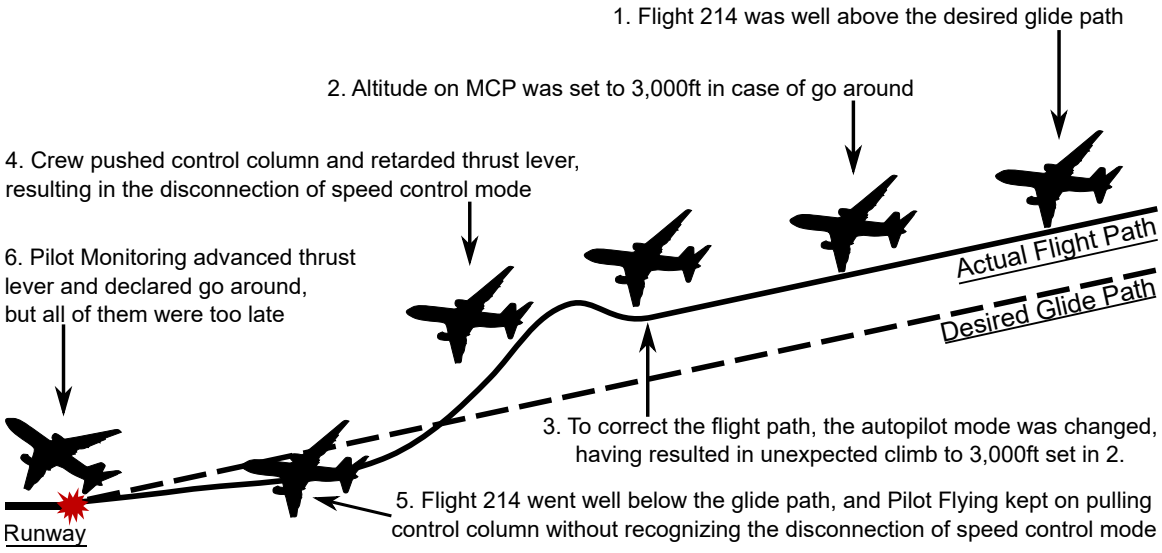


Fig. 2.2: Profile of accident: crash of Asiana Airlines Flight 214 at San Francisco in 2013.

Successive two crashes of Boeing 737 MAX (KNKT, 2019; AIB, 2019)

This is one of the latest cases of such complex accidents occurred in the 2010s. Boeing 737 MAX operated as Lion Air Flight 610 and Ethiopian Airlines Flight 302 doomed soon after the take-off and crashed into the surface in quick succession, i.e., within five months (KNKT, 2019; AIB, 2019). The 737 MAX is the latest model of Boeing 737 series, whose commercial operation started in 2017. The aircraft is designed to improve the efficiency with larger two engines, and they are equipped at further forward and higher position than the conventional models because of their size. At the same time, this caused the pitch up attitude of the aircraft during the flight, and the Boeing implemented an automated system called the Maneuvering Characteristics Augmentation System (MCAS) to prevent the phenomenon. However, there was not enough information about MCAS, provided by the manufacture. One of the main reasons for this was that the MCAS could prevent its sales; one of the advantages of the 737 MAX was that pilots of other 737 series does not require much training to operate this new aircraft. In addition, the manufacture was rushed to construct the 737 MAX due to business competitions in the aviation industry. In this context, the MCAS was suddenly activated and pushed the nose down in both cases; the "fight" between the pilots and MCAS can be observed in the data provided by the Flight Data Recorder (FDR).

These accidents tell us that not only the human and technical factors, but also the impact of organization factors and working environment in a specific situation should be taken into account to describe and understand the safety of highly sophisticated — complex systems today. Systems described from this perspective are so-called socio-technical systems, and Rasmussen and Svedung (2000) proposed a hierarchical model of the system as shown in Fig. 2.3, where relationships between individual workers, teams, and even larger organizations are described.

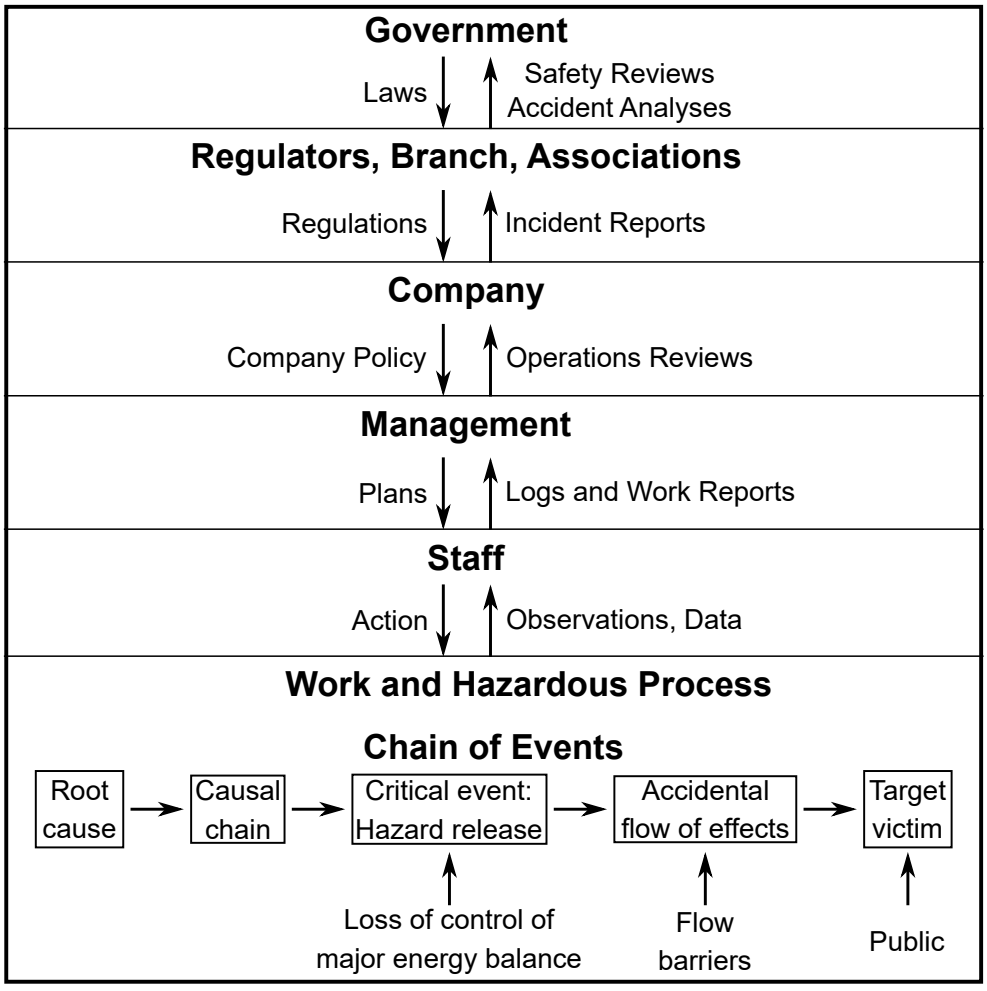


Fig. 2.3: Schematic view of socio-technical systems originally proposed by Rasmussen and Svedung (2000) and summarized by Leveson (2004).

The socio-technical systems and the fields of practice described in Fig. 1.1 can now be regarded as equivalent, and the objective of the resilience engineering is, in short, to manage their evolutions. However, the behavior of the systems is generally "non-linear" in the sense that they are intractable with traditional approaches based on cause-effect relationships. This is a reason why we are confused by the envisioned

world problem, and no effective ways to investigate them have been established yet. In this respect, the next sections reviews historical development of accident models, or how people have been trying to elucidate the nature of accidents.

2.3 Historical Development of Accident Models: Analogical Concepts to Describe Accidents

2.3.1 Sequential Accident Model

Accident models have been developed over almost 90 years, those of which were initially based on the idea of "linear" cause-effect relationships; those models regard a final consequence (e.g., accident) as a result of a chain of preceding events, and the analysis of the chain provides root causes of the consequence. That is why accident models in the early time of their history are called sequential accident model; Domino Model (Heinrich, 1931) shown in Fig. 2.4 is one of the best-known examples. The sequential accident models looks so clear and intuitive to people that they have been widely adopted and remaining as the industry standards towards the end of 1970s (Hollnagel and Goteman, 2004).

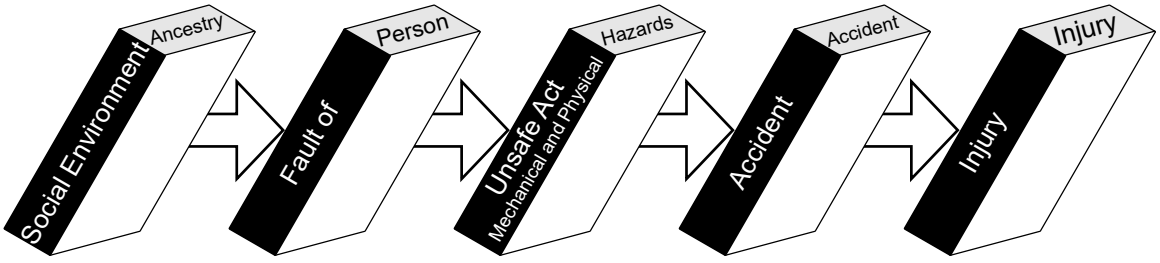


Fig. 2.4: Schematic illustration of domino model.

The situation has changed around the end of 1970s because people became to know that a number of major industrial accidents cannot adequately be explained based on such simple cause-effect relationships. In the early time, human error was assumed to stem from inherent deficiencies and the fact that humans naturally fail to perform tasks, just the same as machines or structures can fail. However, the lessons learnt from major accidents such as the TMI disaster convinced people that the contextual conditions under which a work is performed greatly influence on human performance failure. This resulted in a demand to improve the explanations of human performance under a certain working context and triggered the development of a new accident model.

2.3.2 Epidemiological Accident Model

Epidemiological accident models represent accidents in analogy with the spread of the disease. That is, outcomes are caused by a certain combination of multiple factors, which can be both latent and manifest, existing in a same space and time. This analogy provides an insight that accidents are caused by a certain combination of "agents" and environmental factors existing together at the same time. This enables us to describe a complexity of events in which even if each single factor is too trivial to cause serious outcomes, the effect of their combination could be significant enough to bring about accidents.

Swiss Cheese model (Reason, 2000; Reason et al., 2001) is one of the best known examples of the epidemiological accident model. In this model, there are factors which is called hazards, and they are usually defended by multiple layers represented as Swiss Cheese as shown in Fig. 2.5. The hazards have potentials to cause unwanted outcomes such as accidents, and they will be salient when they break through the all of Swiss Cheese defenses. On the other hand, while each Swiss Cheese defense has enough ability to defend hazards, they also have potential/latent failures represented as holes of the Swiss Cheese. This is because the Swiss Cheese defenses generally involve some functional deficiencies, and it is difficult to completely remove all of them. Therefore, although the undesired outcomes are usually prevented by the layers of Swiss Cheese defenses, the holes of all the Swiss Cheese layers could align under a specific combination of environmental factors, and the alignment enables the hazards to break through the defenses; the impact of the hazards gets salient and results in serious outcomes in the end.

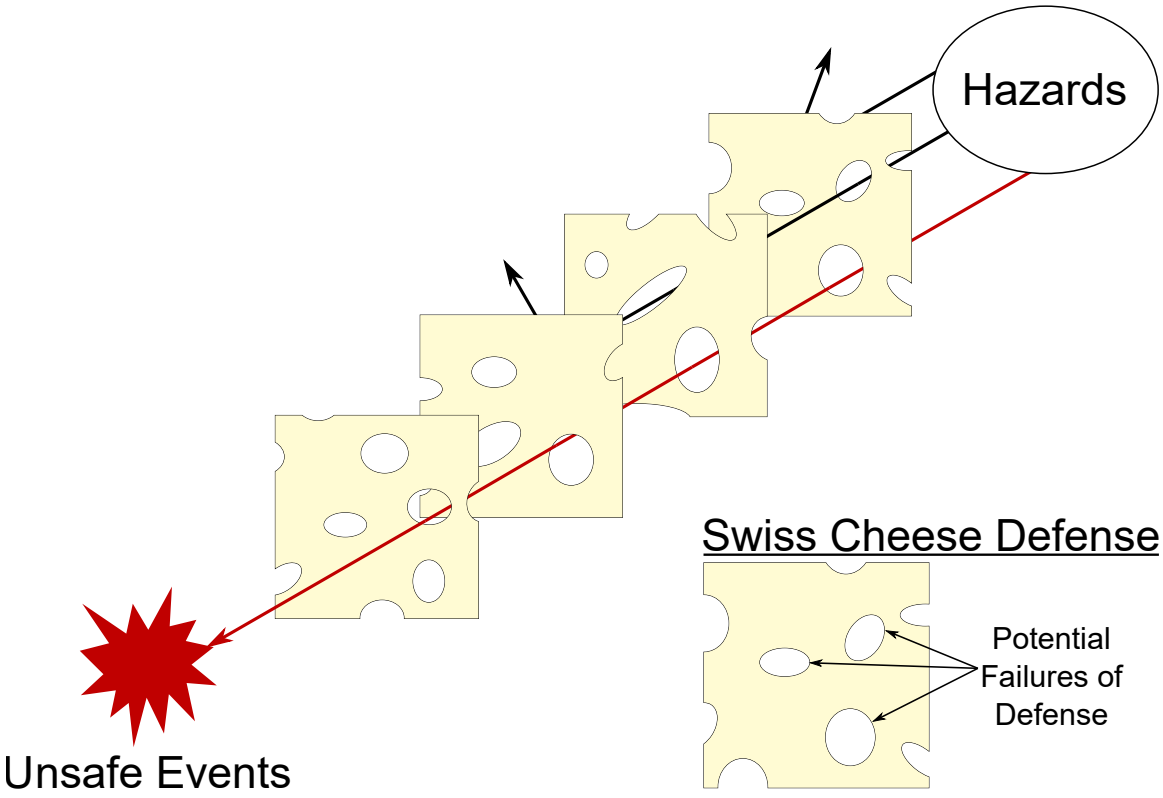


Fig. 2.5: Swiss Cheese model.

The epidemiological models overcame the limitations of sequential models in the sense that they regard accidents as a result of more complex interactions among different-multiple factors. However, this model cannot provide more than concept of how the accidents occur, and it is difficult to step into further details for describing the nature of accidents and their countermeasures. This challenge is exactly what we are currently facing, and a further extended model — systemic accident model has been proposed in response to the limitations.

2.3.3 Systemic Accident Model

Systemic accident models have been developed to describe and understand the safety of socio-technical systems. Its focus is put on characteristic performance on the level of system as a whole rather than that of microscopic mechanism of each component of the systems (Hollnagel and Goteman, 2004). Specifically, the systemic accident model deliberately avoids describing a final consequence as a result of sequential or ordered relationships among events or components making up of a system. Instead, this model tries to represent systems' behavior as a phenomenon emerging out of

the "non-linear" interaction among factors/components of the system. There are currently two major models based on this idea.

The one is Systems-Theoretic Accident Model and Processes (STAMP) proposed by Leveson (2004). In the STAMP, systems are regarded as interrelated components kept in a dynamic equilibrium state by feedback loops of information and control. In addition, the safety of the systems is regarded as a continuous control task of the systems to impose the constraints that are necessary to limit the systems' behavior and maintain the equilibrium; accidents in the STAMP are therefore analyzed in terms of why safety controls/constraints were inadequate, or if they were potentially adequate, why the systems were unable to exert appropriate control over their enforcement. This way of analysis enables us to see a process of accident as an adaptive feedback function that fails to maintain appropriate performance changes over time to cope with a complex set of goals and values. Moreover, the STAMP allows us to investigate non-linear relationships representing the dynamic behavior of the entire technical and organizational structure over time.

The other one is Functional Resonance Accident Model (FRAM) (Hollnagel, 2004) whose schematic illustration is shown in Fig. 2.6. In this model, accidents are regarded as emergent phenomena resulting from non-linear interactions among variabilities inherent to human, machines, and their surrounding environment. According to the concept, operators of the socio-technical systems are generally demanded to precisely carry out Standard Operational Procedures (SOPs) that adhere to the instructions or rules issued by organizations (e.g., the company or the government). However, in reality, it is quite difficult in many cases because there may exist variabilities of the working environment caused by temporal conditions such as available resources (e.g., time) or the existence of simultaneous goals to be attained, and they interfere with execution of the SOPs. These conditions tend to create a trade-off in the operations of socio-technical systems; although operators are fundamentally required to execute predetermined procedures precisely, they should perform them in a more flexible way to cope with the situation (e.g., deviations from the SOPs). This dilemma is referred to as the Efficiency-Thoroughness Trade-Off (ETTO) principle, and in most cases, the operations of socio-technical systems cannot escape from it. In other words, operations are always having to adapt to a given situation, and such adaptations generate functional variabilities of task performance in humans/machines. Those variabilities consequently interact with each other, and in a specific context, some of them that are usually too weak to notice could go beyond our noticeable/acceptable threshold as if they were resonated. This is referred to as the functional resonance, and the

FRAM enables us to describe and understand the accident of socio-technical systems emerging out of such non-linear effect of the variabilities.

By the way, the Functional Resonance Accident Model is now updated and renamed as Functional Resonance "Analysis Method" (FRAM) (Hollnagel, 2012b). This is because the viewpoint of the FRAM is no longer limited to only accidents or why things went wrong; the FRAM additionally addresses normal daily activities and why things go right. In other words, the FRAM had to be updated from the perspective of Human Reliability Analysis (HRA), and this demand consequently accelerated the development of the resilience engineering. The next section reviews the historical development of the HRA methodologies and summarizes what was still missing in the traditional approaches.

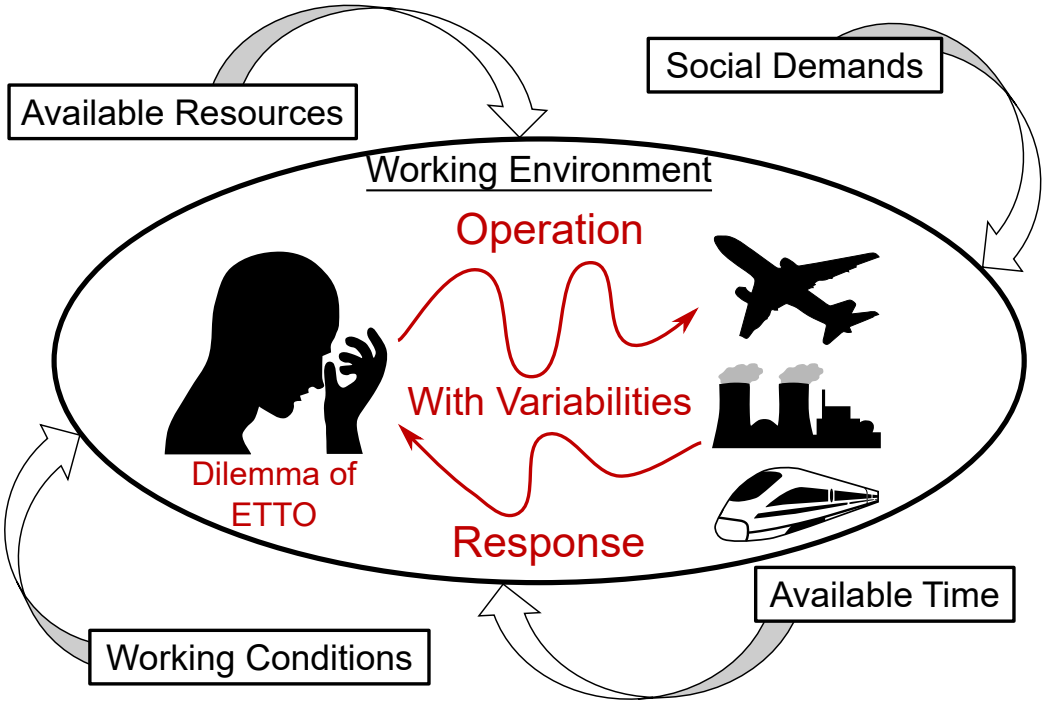


Fig. 2.6: Functional Resonance Accident Model.

2.4 Historical Development of Human Reliability Analysis

2.4.1 Origin of Human Reliability Analysis

When technical factors were playing significant roles in the safety of artifacts, mechanical reliability had been crucial interest of the people and actively investigated; the typical approach was to prepare the probability of errors that each mechanical

component could cause and calculate their sequential combinations. However, it was found that the reliability of human beings also has significant effect on the safety as well as, or even more than mechanical reliability, and the demand to investigate human reliability became louder and louder.

Human Reliability Analysis (HRA) has been developed in this context. The HRA is to investigate the reliability of human beings in terms of deviations from expected safe behavior. The deviation is generally caused by combinations of multiple factors such as difficulty level of tasks, their sequential order, and the effect of their surrounding context. In this regard, the HRA enables us to estimate the probability of human to perform unsafe actions, depending on the times; their methodologies can be classified into the following — first and second generations.

2.4.2 First-Generation Human Reliability Analysis

The first-generation HRA has initially been developed for the Probabilistic Safety Assessment (PSA) of nuclear power plant. There were few early developments of the methods in the last half of 1970s. Then, most of them were developed in the first half of the 1980s, and its enthusiasm declined after that; some methods were kept on developing in 1990s as well, but the scale was much smaller than it used to be in 1980s. It should be noted that this trend is clearly related to the TMI disaster occurred in 1979 (Hollnagel, 1998).

The basic approach is to prepare the probability of tasks carried out by human (e.g., turning on a switch or connecting to power) to fail, and calculate their sequential combinations to estimate Human Error Probability (HEP) of the entire sequence. For this purpose, several methods, as shown below for example, have been developed.

Accident Initiation and Progression Analysis (AIPA) (Raabe et al., 1976)

The AIPA has been developed as a method to estimate the reliability of operators' responses in operations of large high temperature gas-cooled reactors. The operator's actions were regarded as bimodal, i.e., success and failure, which was defined as whether required processes are made. The purpose of AIPA was to evaluate the probability of whether the required actions are carried out, described in terms of the mean time to operator response. The AIPA method was used to estimate the probabilities for each action, based on expert judgement.

The problem with the AIPA is that operators are basically seen as a black box that emits a possibly successful response according to "unknown mechanism." That

is why it remains unclear that how their responses are actually brought about. In addition, status of the operators are described by only success and failure in this method; although the AIPA very clearly exhibits the fundamental features of the first-generation HRA, it is still unsophisticated (Hollnagel, 1998). It should be noted that the limitation of the AIPA might be taken for granted because the AIPA is one of the very initial methods of the first-generation HRA developed in the early 1970s.

Technique for Human Error-Rate Prediction (THERP) (Swain and Guttmann, 1983)

The THERP is one of the best-known methods of the first-generation HRA. The purpose of the THERP is to calculate the successful performance of actions required to complete tasks. The THERP is carried out by using a event tree diagram representing a series of tasks and pre-determined HEP for each of them as shown in Fig. 2.7. Also, the HEP is modified to consider the effect of Performance Shaping Factors (PSFs) including dependency among operators, stress levels, or experience, for example. Based on this idea, the THERP consists of the following six steps:

1. Define the system failures caused by human errors and for which probabilities are to be estimated.
2. Analyze human operations and identify relationships among them.
3. Estimate the relevant human error probabilities.
4. Determine the effects of the human errors on the system failure events.
5. Find out and suggest countermeasures in order to reduce systems failure rate to an acceptable level.
6. Review the consequences in terms of availability, reliability, and cost-benefit.

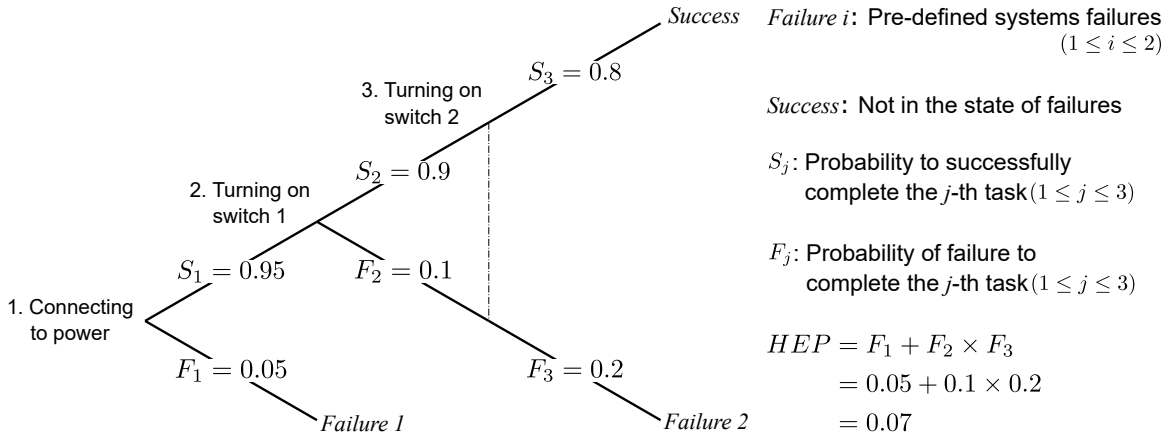


Fig. 2.7: Event tree diagram of THERP: example (Sawaragi et al., 2018).

The THERP has been regarded as one of the few and most sophisticated methods of the first-generation HRA. This is mainly because it well describes how events, moreover, human actions should be modelled and quantified. On the other hand, there still remains room for improvement since the event tree still can describe only binary choices, i.e., success and failure (Hollnagel, 1998). Therefore, it is difficult for this method to deal with more complex problems.

Cognitive Environment Simulation (CES) (Woods et al., 1988)

The CES can be regarded as a method of the first-generation HRA, but it is somewhat different from other typical approaches. The characteristic of the CES is that it generates possible errors via a simulation, instead of theoretically defining them like other methods and investigates how an operator will respond in a situation. Moreover, the CES simulates an intention to act, or more specifically, actions that an operator is likely to take in different situations by using time sequential data representing a plant state. This intention to act is then sent back to the simulation process, and a new set of the time sequential data is generated. The process is repeated, and the possibility of incorrect actions is introduced; it should be noted that the actions of operators are no longer regarded as bimodal, i.e., success and failure.

The development of CES was a unique approach as an alternative to the existing first-generation HRA methods. However, it depends on very detailed operators model supported by artificial intelligence. This is critical because the performance of computers at this time was so limited that the CES was far from practical.

The first-generation HRA methods had been developing with involving some limitations in this way. Especially, one of the biggest problems with the methods is

that they are still remaining in a framework in which human error is caused by their inherent deficiencies, and human beings naturally fail to perform tasks, just the same as machines or structures can fail. However, the human beings are far more complex than described in the methods of the first-generation HRA, and it is not realistic to estimate the reliability with simple linear summation of components. The second-generation HRA has been developed to overcome this problem.

2.4.3 Second Generation Human Reliability Analysis

Through the experience and lessons of the first-generation HRA, the interest to estimate human reliability stemming from their complexity became louder and louder. Specifically, the importance of contextual conditions under which tasks are performed attracted the attention. This trend has changed the viewpoint of human error; the human error is no longer a cause to threaten the safety but should be regarded as a result caused by the effect of surrounding environment, i.e., Error-Forcing Contexts (EFCs). The second-generation HRA therefore put focuses on the effect of EFCs that make the HEP extremely high under a specific combination of them; it should be noted that EFCs are usually provided as verbal descriptions.

A Technique for Human Event Analysis (ATHEANA) is a representative method of the second-generation HRA (Barriere et al., 2000). Bearing in mind that the environment and the surrounding context may affect the behavior of a human operator, the ATHEANA takes account of the EFCs, which are then combined with the PSFs. This method can analyze the occurrence of an actual unacceptable event and clarify how to improve safety. The ATHEANA used to be classified as the first-generation HRA in 1990s (Hollnagel, 1998) but is now regarded as the second-generation HRA in 2010s (Sawaragi et al., 2018).

Cognitive Reliability and Error Analysis Method (CREAM) is also a representative method of the second-generation HRA (Hollnagel, 1998). In the CREAM method, Hollnagel (1998) referred to contextual conditions collectively as Common Performance Conditions (CPCs). Then, CPCs are used to identify an index which is called control mode representing the state of EFCs. The CREAM is, therefore, can be regarded as an extended method of the ATHEANA.

The CREAM has been applied to investigate human reliability in many fields so far. Akyuz and Celik (2015) adopted CREAM to assess human reliability along with the cargo loading process onboard liquefied petroleum gas (LPG) tanker ships. They developed a human error assessment approach based on CREAM and applied that to the monitoring of the crew's cognitive actions or attitudes during cargo operations

onboard LPG tankers. They concluded that the approach could be applied to any other critical operational processes. Zhou et al. (2017a) evaluated the human reliability of seafarers performing their onboard operations. In their work, they introduced eight customized CPCs for better capturing the essential aspects of the working situations and conditions for tankers. They also built a model using the Markov method to estimate a quantified human error probability (HEP) and identified the result of the error probability intervals are limited within the tolerant ranges using the original CREAM. Zhou et al. (2017b) incorporated CREAM and Monte Carlo simulation into Fault Tree Analysis (FTA) (Lee et al., 1985) to evaluate a Liquefied Natural Gas (LNG) carrier spill accident. They constructed a modified FTA model for LNG spill accidents during LNG carriers' handling operations and introduced the CREAM model to predict human errors in the operations. Their results were synthesized in the end so that Monte Carlo simulation can provide risk as intervals of probability. Moreover, in the nuclear energy field, Yoshida et al. (2002) applied CREAM to evaluate the effectiveness of Accident Management (AM) which is prepared for unexpected emergencies based on Probabilistic Safety Assessment (PSA). They developed a new method to quantify the decision-making failure probability of an emergency organization facing an emergency of nuclear power plant, which provides an AM strategy, by using THERP and CREAM. Then they applied it to the case of a typical Pressurized Water Reactor (PWR) plant. In conclusion, they found their method can work effectively even if the analyst is not a professional of human reliability engineering field and is applicable to other fields.

Despite the great progress from the first-generation HRA, the second-generation HRA also involves some limitations. Hollnagel (2012a) points out this more than 10 years after the development of the CREAM as follows on his website:

> Although CREAM still appears to be used and referenced, it is only fair to point out that the method from my point of view is obsolete. There are several reasons for that. First, because it focuses on how actions can fail, rather than on the variability of performance, i.e., a Safety-I perspective (q.v.). Second, because it focuses on one part or 'component' of the system only, namely the human(s). While this seemed sensible in the aftermath of the debates around first and second generation HRA, it can now be seen as representing a structural rather than a functional viewpoint. Third, because it indirectly lends support to the concept of 'error'. (It doesn't really, of course, but no one seems to have noticed that.)

> Seen in hindsight, only the A (for Analysis) and M (for Method) make sense. Cognitive Reliability (the 'CR') is a misleading oversimplification: explaining human performance as based on 'cognitive processes' represents a myopic information processing view, and talking about the reliability of such processes is an artefact of the PRA/PSA mindset. 'Error' is, of course, theoretically vacuous as explained elsewhere. (Not that this has any effect on its popularity.) So CREAM really ought to have been called CVEAM ('V' for Variability), or CVAM (getting rid of the 'E'), or perhaps just VAM. But then it is only a short step to FRAM, which in a way is what CREAM could have been if I had known then what I know now. But I didn't.

In short, the second-generation HRA is unintentionally used to investigate how actions can fail in many cases, and this viewpoint is no longer effective in the current context of the safety management. In addition, the second-generation HRA intrinsically focuses on a limited part or components of a target system, i.e., humans only, and its representation is structural rather than functional. That is why the second-generation HRA is now obsolete, and new approaches which replace the traditional HRA methodologies and overcome their limitations are required.

2.5 Paradigm Shift from Safety-I to Safety-II: Development of Resilience Engineering

2.5.1 Safety-I and Safety-II: New Perspective of Safety

Across all industries and professions, safety has been associated with being free from harms, where no unacceptable events such as failure, damage, or accidents occur. The focus of traditional safety management is therefore put on how to identify and eliminate potential harms, or why things go wrong as much as possible in advance. This approach is consistent with the well-known definition of the safety issued by ICAO (2009) as well. However, people came to realize that it is inadequate for the safety of socio-technical systems today, and an alternative approach has to be developed; it is to identify harm-free everyday works or why things successfully go right and facilitate lessons learnt from such positive aspects of the safety. Hollnagel et al. (2013) has referred to the two approaches as Safety-I and Safety-II, respectively, and advocated a paradigm shift from Safety-I to Safety-II.

Hollnagel (2017) points out that the most serious problem with the traditional approach for the safety is that although it makes intuitive sense to be free from incidents and accidents, it does not make much sense that the goal of safety management is to be without something. According to him, an increase in the safety is currently represented by a decrease in what is measured; the lower number of reported accidents or other unwanted outcomes is regarded as a higher level of safety, and safety management has been carried out based on this point of view, i.e., Safety-I approach. However, this kind of approach limits our opportunities to know how well the safety management works only when something undesirable happen; the better the safety management works, the less information about how to make improvements there is. This irony is generally known as the "regulator paradox" (Weinberg and Weinberg, 1979) as well, suggesting that the absence of feedback ultimately leads to a loss of control. Therefore, the traditional safety management approaches, i.e., Safety-I intrinsically involve a critical problem that although the goal of traditional safety management is to be without something unacceptable, the achievement of the goal could paradoxically lead to the loss of the safety.

The Safety-II approach has been developed in response to this problem. The safety now requires to be described from the perspective of its "existence," rather than absence of unacceptable events, otherwise opportunities to obtain the feedback or control for the safety will be lost. In addition, the frequency of unacceptable events is no longer high today; it is instead filled with a number of normal daily activities and successful good practices, according to Fig. 2.8. The focus of the safety management should, therefore, be put on why things go right rather than why things go wrong and fail. This idea resulted in the paradigm shift from Safety-I to Safety-II, which further led to the development of resilience engineering. Here, it should be noted that the Safety-II does not neglect to consider why things go wrong; it considers all of the events as a whole, as also shown in Fig. 2.8.

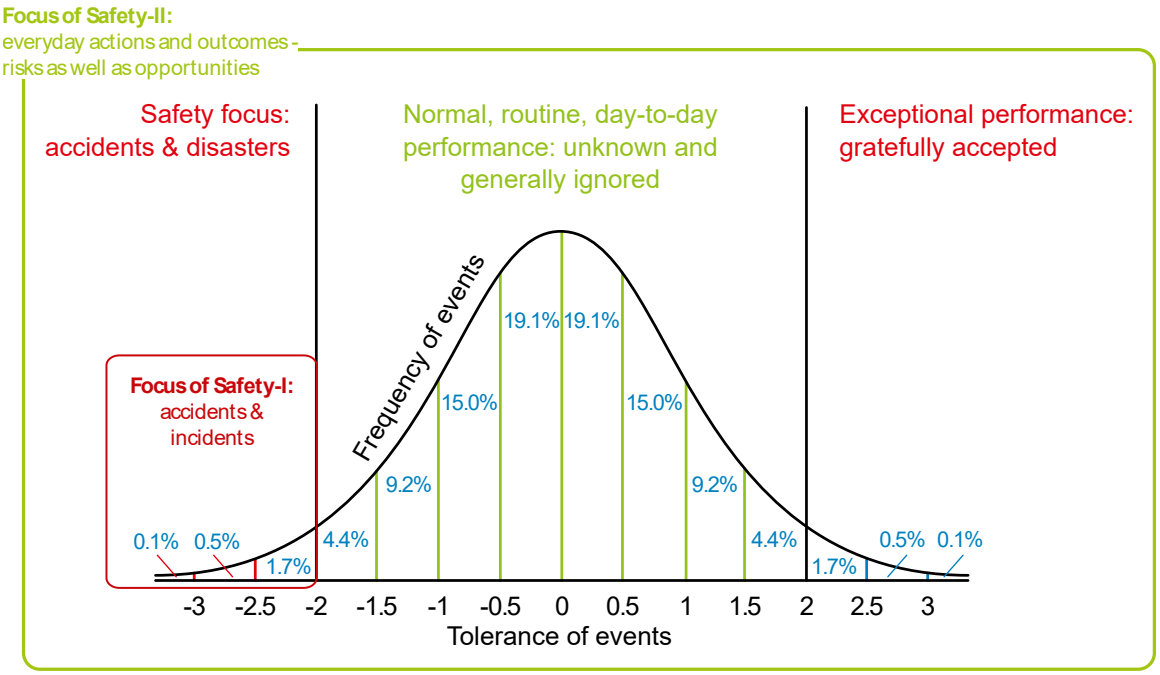


Fig. 2.8: Relationships between tolerance and frequency of events with focus of Safety-I and Safety-II (Hollnagel et al., 2013).

2.5.2 Resilience as Property of Socio-Technical System

The concept of resilience has been spreading from its origin to a variety of fields. The resilience originates from the physics of material science, and its general definition is 1) the ability to become recovered, happy, or prosperous again after a difficult situation or event, or 2) the ability of a substance such as rubber to return to its original shape after it has been deformed or bent, according to the Longman Dictionary of Contemporary English. The resilience is also applied to the field of psychology after a while to address a mental ability of individuals to withstand traumatic stress. In this context, the resilience, which is now considered as a milestone of the resilience engineering, was developed by Holling (1973) and Carpenter et al. (2001).

Holling (1973) has come up with an idea that ecological systems could be described in terms of two properties: stability and resilience. The stability here is regarded as a systems' capacity to return to an equilibrium state after temporal disturbances. On the other hand, the "ecological resilience" is referred to as a systems' ability to absorb changes, identify next compatible states, and get into the states for their survival. The concept of the "ecological resilience" was applied to the field of social science and further extended by Carpenter et al. (2001).

The resilience changes depending on continuous adaptation cycle of a system to its surrounding environment, according to Carpenter et al. (2001). Moreover, the adaptation cycles generally involve feedbacks of entities (e.g., disturbances) between the systems and their surrounding environment, which produces alternate equilibriums or stable states to survive. Therefore, the "ecological resilience" has the following three primary properties, that are often discussed within the framework of Complex Adaptive System (Mitchell, 2009; Johnson, 2009) as well:

- The amount of change a system can undergo while maintaining the ability to function.
- The degree to which a system can organize itself.
- The degree to which a system can develop a capacity to learn and adapt.

The resilience of socio-technical systems was derived from this idea, which is now regarded as "an ability of the systems to function as required under expected and unexpected conditions alike." As socio-technical systems get larger and intractable, people have realized the importance to ensure this ability in addition to the traditional safety management approaches, which inevitably led to the paradigm shift from Safety-I to Safety-II as well. The resilience has been attracting attention especially since the Fukushima Daiichi nuclear disaster in 2011. Also, the ditching on the Hudson river of US Airways Flight 1549 known as the "Miracle on the Hudson" (NTSB, 2010) is often taken as an example for a "resilient success story" in which everyone on board survived even though the flight experienced fatal damage on their both engines. The resilience engineering has been developed on the basis of these experiences.

Besides, Hollnagel (2017) has initially proposed that the ability of adaptation in the "ecological resilience" should be revised for the resilience of socio-technical systems. This is because ecological systems are generally limited to be reactive, and their adaptation processes tend to be passive; socio-technical systems, on the other hand, can consider what may happen, utilize the information, and direct their behavior since they include cognitively active elements, i.e., humans/organizations. That is why the ability of adaptation should be specified by taking into account the cognitive aspect of the socio-technical systems; it is now replaced with the potentials of the systems to monitor, anticipate, respond to, and learn (get feedbacks) from expected/unexpected conditions alike in the resilience engineering. The resilience engineering, in summary, is to investigate and enhance the resilience of socio-technical systems consisting of these potentials, and the Functional Resonance "Analaysis Method" (FRAM) (Hollnagel, 2012b) has been developed for the purpose of it.

In another perspective, the resilience engineering is to validate the feasibility of socio-technical systems evolving under specific contexts. The paraphrased purpose of the resilience engineering is, therefore, to manage the evolution of the socio-technical systems so that they remain in acceptable domains, and their advantageous evolutions will go on as shown in Fig. 2.9. The FRAM enables us to envision such continuous evolutions in Research and Development (R&D) processes. In addition, the socio-technical systems and the fields of practice can be regarded as equivalent as previously described. We can see the intersection of the envisioned world problem, resilience engineering, and FRAM here.

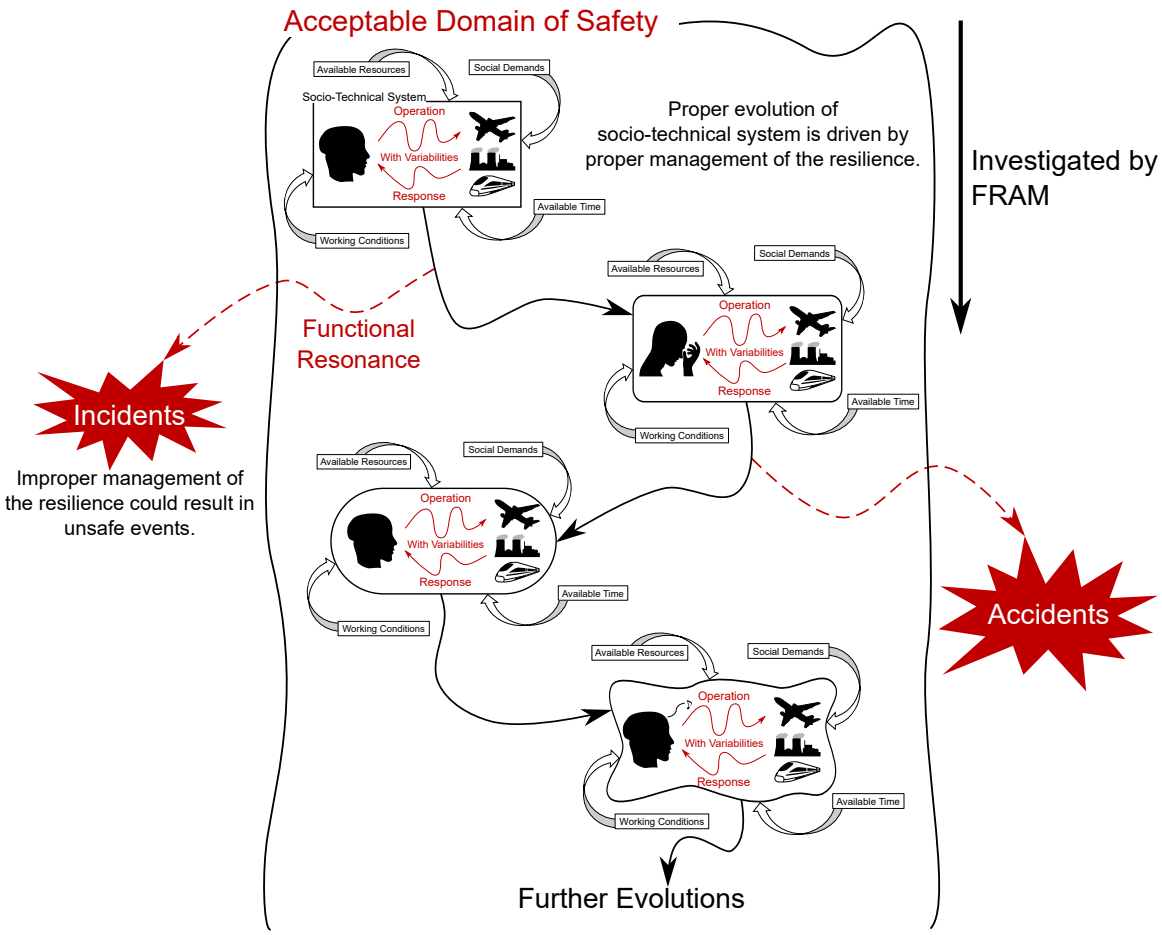


Fig. 2.9: Resilience engineering: proper management of resilience for proper evolution of socio-technical systems.

Chapter 3

Development of Simulation Model Based on Functional Resonance Analysis Method

3.1 FRAM as Proposed Method

The Functional Resonance Analysis Method (FRAM) (Hollnagel, 2012b), which used to be the Functional Resonance Accident Model (Hollnagel, 2004), has been developed in response to the demand of the resilience engineering. It enables us to envision the evolutions of socio-technical systems and provides insights to enhance their resilience with us. Specifically, the FRAM investigates the safety/feasibility of the socio-technical systems emerging out of the non-linear interactions among variabilities existing in their functions.

There have actually been a number of methodologies developed for safety analysis of human-machine systems such as Fault Tree Analysis (FTA) (Vesely et al., 1981; Lee et al., 1985; Stamatelatos et al., 2002), Technique for Human Error-Rate Prediction (THERP) (Swain and Guttmann, 1983), Petri Net (Petri, 1966; Brauer and Reisig, 2009), Structured Analysis and Design Technique (SADT) (Ross, 1977), or Function Analysis System Technique (FAST) (Bytheway, 2007), for example. Although they are ones of the most well-known approaches of safety analysis and have been applied to many fields, their limitations have also been pointed out so far. The FTA and its successor model: THERP are typical sequential methodologies of reliability analysis; the limitation is that their approach is so simple and linear that it is difficult to address complex aspects of the safety such as cognitive behavior of human beings and "non-linear" interactions among the systems' components. The SADT and FAST are good at describing functional relationships within the systems as a diagram, but its

expressiveness (i.e., semantics of functions) is still limited compared to FRAM. The Petri Net is also a representative approach of discrete event systems, but it is also limited to deal with the "non-linear" aspects of the safety as well as other approaches.

The FRAM is characteristic compared to these well-known methods in terms of the following aspects. First, the scope of FRAM is not limited to failures or "why things go wrong"; its aim is rather to find out "why things go right" and facilitate the safety based on the findings as its fundamental concepts, i.e., Safety-II and Resilience Engineering suggests (Hollnagel, 2012b, 2017). Also, the FRAM is not to investigate just "snapshots" of safety or systems' state based on linear causality, contrary to the traditional approaches; it has a potential to envision dynamic characteristics of the safety or systems' state emerging out of "non-linear" interactions among components of the target systems. Moreover, such emergent behavior essentially involves abstract, ambiguous, and qualitative aspects, for which the balance between qualitative and quantitative perspective must be taken into account; the FRAM could provide qualitative comprehensions of the emergent behavior on the basis of quantitative criteria. The FRAM is different from other traditional approaches and adequate for investigating the safety of complex human-machine interactions in these respects.

The FRAM is based on the following four principles describing essential roles in the safety of socio-technical systems:

3.1.1 Four Principles

The first principle is the equivalence of success and failure. Conventionally, the state of a system is thought to be bimodal: function or malfunction, as shown in Fig. 3.1. This idea originated in a time when artifacts were simpler and thought to be more independent than they are currently. Therefore, sources of success and failure are completely independent of each other, and eliminating the sources of failure (such as malfunction of machines or human errors) is assumed to enhance the safety accordingly, which is also consistent with the idea of Safety-I.

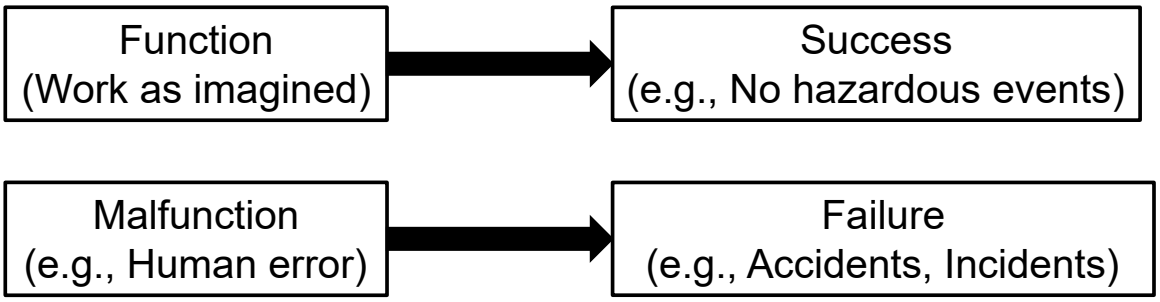


Fig. 3.1: Traditional idea of success and failure: they stem from different sources.

However, the situation has been changing due to the increased complexity of both mechanical systems and society. Specifically, system states can be multimodal in the sense that they are variable and flexible between "function" and "malfunction." In other words, success and failure are equivalent because, depending on specific contexts, they possibly come from the same source, as shown in Fig. 3.2. Therefore, it is no longer feasible to consider only success or failure individually, nor to seek just one "root cause" of accidents or success stories.

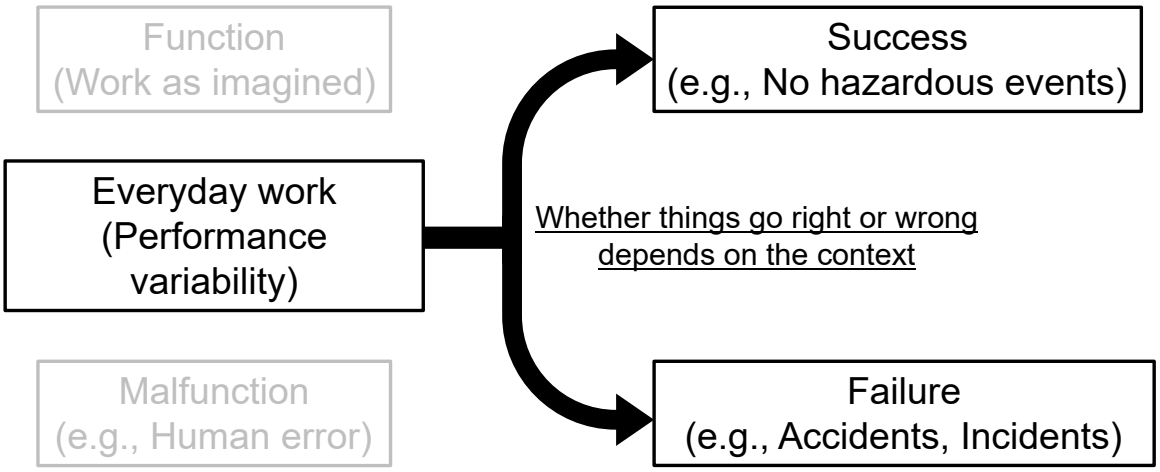


Fig. 3.2: Equivalence of success and failure: the same factors could lead to various outcomes depending on contexts.

The second principle corresponds to the ETTO principle, which is also referred to as the approximate adjustments. This principle was prepared for describing how variabilities can be generated in socio-technical systems. As already described in the ETTO principle, operators of the socio-technical systems need to adapt their works/tasks to their actually facing situations contrary to expected — ideal situations as shown in Fig. 3.3. This creates a gap between what is called work-as-imagined and work-as-done and results in functional variabilities of humans/machines.

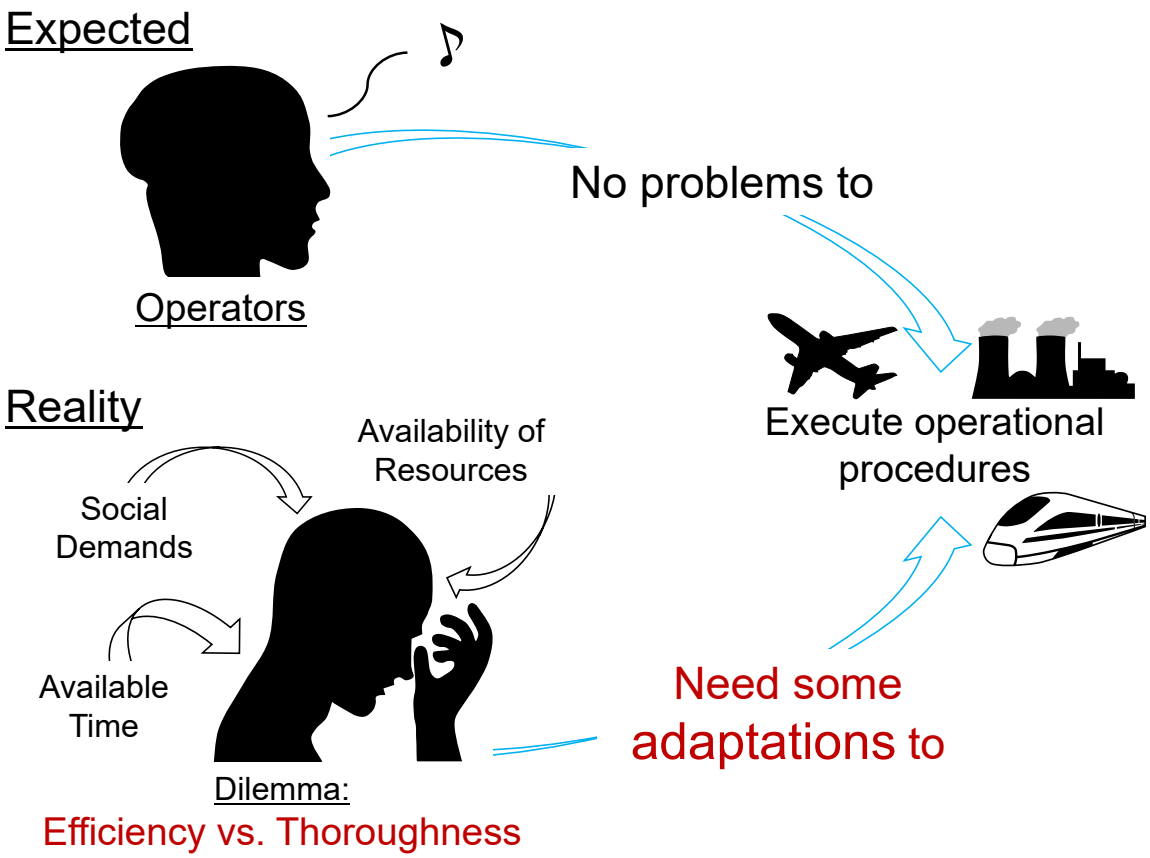


Fig. 3.3: Approximate adjustments: how the difference between work-as-imagined and work-as-done is caused.

The third principle is the emergence. Operations of socio-technical systems often face unexpected outcomes, most of which are difficult to explain with decomposition and causality — a typical way to investigate what has happened. In such cases, the outcome is said to be emergent rather than resultant; we regard something as "resultant" if it is tractable based on known processes or developments. Here, note that emergence does not mean that something happens "magically"; it means that something happens whose process cannot be explained by cause-effect relationships.

The fourth principle is the functional resonance, which is introduced to support the third principle, emergence. Traditionally, "resonance" refers to phenomena in which a system oscillates with large amplitude when the oscillating components come together with specific frequencies. For example, random noise could make weak signals exceed the detection threshold with the principle of resonance, known as stochastic resonance, which is usually used to enhance the sensitivity of a device. The FRAM uses this physical phenomenon as an analogy to clarify how the interaction among variabilities can lead to emergent phenomena. Specifically, the functional resonance represents

phenomena in which functional variabilities that are usually too weak to notice could go beyond acceptable threshold as if they resonated in a specific context.

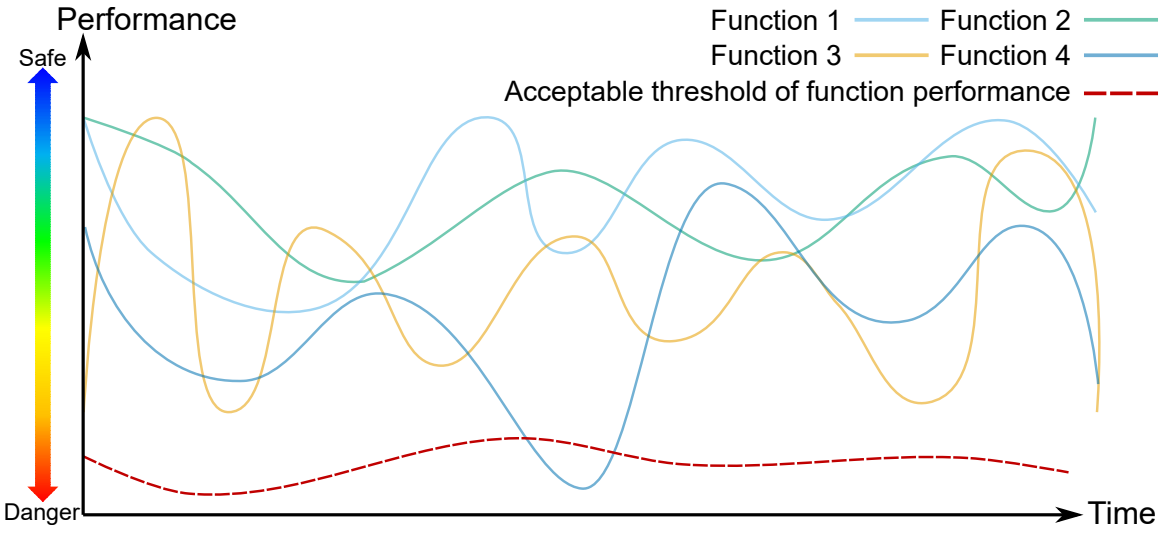


Fig. 3.4: Schematic illustration of functional resonance.

The role of the four principles in the safety of socio-technical systems can be summarized as follows: it is no longer feasible to seek just one "root cause" of accidents or success stories because success and failure are equivalent rather than independent. Instead, the safety of the socio-technical systems depends highly on functional variabilities resulting from approximate adjustments; the non-linear interactions of the variabilities could cause emergent phenomena driven by the principle of functional resonance in the end. The FRAM investigates the safety of socio-technical systems based on this perspective.

3.1.2 Procedure of FRAM

The first step of the FRAM is to identify functions of a target socio-technical system. Functions in FRAM are defined as what has to be done to achieve a specific goal such as each task described in manuals or procedures, for example. Also, the functions are defined with six aspects listed in Table 3.1. The functions are connected through these aspects, and a model of socio-technical system is represented as a network of them in FRAM. If the purpose of FRAM is to investigate anomalous events, functions can be obtained from databases related to those events, such as accident reports. Also, if the purpose of FRAM is to simulate the effect of variabilities with an "If-Then Exercise," various methodologies such as Hierarchical Task Analysis (HTA) (Kirwan and Ainsworth, 1992) are available.

Table 3.1: Six aspects of functions.

Aspect	Description
Input	Input/Trigger of functions
Output	Outcome of functions
Precondition	Conditions that must be satisfied before functions start their process
Resource	What is consumed by functions (e.g., fuel, energy, labor force)
Control	What supervises or restricts functions
Time	Time required for the process of functions

The second step is to identify potential coupling among them. Potential coupling refers to the dependencies that can exist among functions. An easy way to find these couplings is to consider the linguistic relationships between the output of one function and other five aspects (i.e., input, precondition, resource, control, time) of the rest of identified functions. For example, suppose that we have three functions: To start car, To release foot brake, and To shift from Park to Drive. The function To start car is triggered by releasing the foot brake, and the gear must be shifted from park to drive before the departure. Therefore, the input of To start car can be "Foot brake is released," and the precondition of this function can be "The gear has already been shifted from park to drive" as well. On the other hand, the output of To release foot brake and To shift from Park to Drive can inevitably be "Foot brake is released" and "The gear has already been shifted from park to drive," respectively. These outputs correspond to the input and precondition of To start car, building a potential network of functions, as shown in Fig. 3.5; the potential coupling is represented as dotted line mong the functions. The manner of finding potential couplings is currently qualitative rather than quantitative, and a more systematic way is desired (Hollnagel, 2012b).

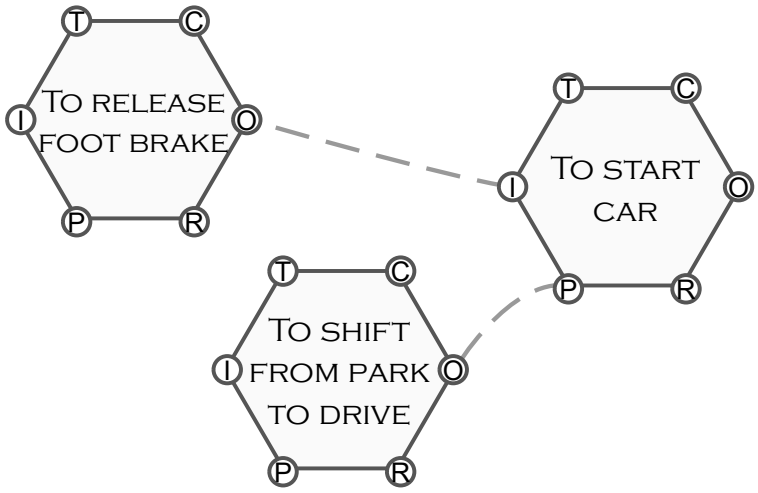


Fig. 3.5: Example of potential coupling: starting a car.

The third step is to instantiate the identified potential coupling. In this process, a network with a specific pattern of dependencies among functions, called an instance, is instantiated depending on situations, or specifically, variabilities as shown in Fig. 3.6. For example, a predetermined procedure or "work-as-imagined" can be defined as an initial instance, and it can be further instantiated by subsequent situations or additional variabilities. procedures that emerge as "work-as-done" are other possible instances. The instantiations provide various "actual works" as instances, and the transition of their state, including their physical structures and safety affected by variabilities, consequently represents the evolution of socio-technical fields of practice.

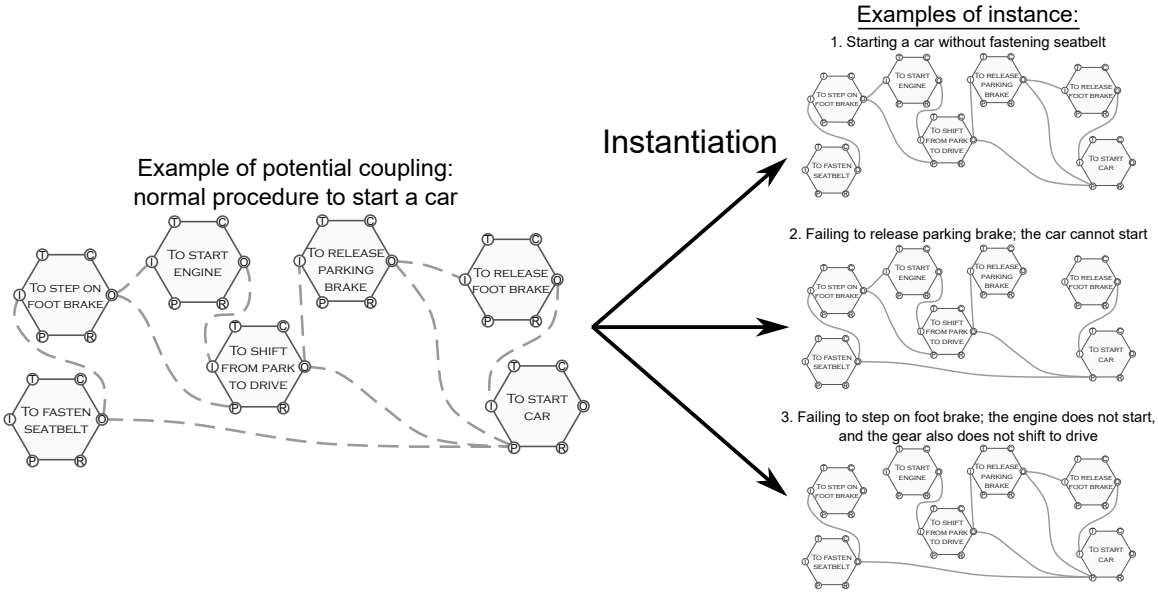


Fig. 3.6: Example of instantiation: normal procedure to start a car.

The fourth — last step is to analyze the safety or feasibility of the instances: some improvements should be made if the result suggests the lack of resilience, otherwise the reason why it is successful should be aggregated. This process is basically qualitative and requires repetitive trials and errors. FRAM investigates the effect of variabilities existing in a specific instance in this way.

3.1.3 Challenges to Utilize

FRAM has been proposed as a method rather than a model (Hollnagel, 2012b). This means that the original FRAM just provides the concept of how to describe and understand the safety of socio-technical systems, and it is necessary to implement a specific model for its practical use. However, this is not easy especially because the definitions of essential entities of FRAM — such as variabilities, propagations, and their interactions are still too ambiguous to deal with. This is one of the major reasons why applications of the FRAM often looks too theoretical and elusive.

To this end, we have been developing FRAM models by introducing numerical definitions of those entities. The feature of those models is that the variabilities of socio-technical systems are defined based on Fuzzy CREAM — an extended model of Cognitive Reliability and Error Analysis Method (CREAM) (Hollnagel, 1998), and the interactions among those variabilities are formulated as well.

There have already been extensive researches devoted to the construction of quantitative FRAM models. Duan et al. (2015) integrated FRAM with a computer tool, model checking (Clarke et al., 1987), to define the interaction of variabilities among FRAM functions so that it can automatically search the potential paths that could lead to hazards. Yang et al. (2017) extended this model by adopting the Simple Promela Interpreter (SPIN) (Holzmann, 2004), which is a kind of model checking, to illustrate functional resonance or emergence. Patriarca et al. (2017) proposed a semi-quantitative FRAM model based on Monte Carlo simulation that can highlight critical FRAM functions and the critical links among them. Slater (2017) represented a network of FRAM functions as a Bayesian network to evaluate the state of each function with probabilistic values. Lee and Chung (2018) tried to quantify the effect of variabilities existing in human-system interaction and proposed a model comprising a FRAM instance and a network of operators based on the heterogeneous network theory to suggest the critical part of human-system interaction and support the management of those variabilities.

Compared with the above approaches, our FRAM model is characterized as an approach that provides qualitative comprehension of the safety of socio-technical systems on the basis of quantitative criteria. Any discussion of the safety of socio-technical systems is apt to contain some ambiguity in terms of evaluating the safety, and it would be preferable to develop a new method that enables evaluation both from the qualitative and quantitative perspectives. However, the FRAM models proposed so far tend to focus on only the qualitative or quantitative aspect, not both. Moreover, these models currently put focuses on numerical evaluation of variabilities in each FRAM function, and systemic behavior of socio-technical systems has not been addressed well yet. Our proposed FRAM model takes this point of view into account: specifically, the FRAM model can support the qualitative interpretation of the safety of target systems based on quantitative properties of each FRAM function.

The following three sections describe the way to develop the FRAM model. There have been two updates of the FRAM model after the development of the initial model; each of the three section presents each FRAM model, respectively.

3.2 Initial Model: Development of Primary Mechanism of FRAM Functions

3.2.1 Numerical Definitions of Variabilities with Fuzzy CREAM

To begin with, numerical definitions of variabilities are required, and Fuzzy CREAM is adopted in this model. The original CREAM (Hollnagel, 1998), which has been presented in the previous chapter, is to investigate how correctly cognitive tasks can be carried out under a certain situation. The Fuzzy CREAM is an advanced model of the CREAM, enabling us to represent variabilities of the working environment and functions quantitatively.

In the CREAM method, Hollnagel (1998) referred to contextual conditions collectively as Common Performance Conditions (CPCs) and then defined and classified them into nine factors: “Adequacy of organization,” “Working conditions,” “Adequacy of man-machine interface,” “Availability of procedures,” “Number of simultaneous goals,” “Available time,” “Circadian rhythm,” “Adequacy of training and experience,” and “Crew collaboration quality.” After that, two more CPCs, “Available resources” and “Quality of communication,” were added to the original nine when FRAM was proposed for the first time (Hollnagel, 2004). Each CPC contains various CPC levels and effects, as shown in Table 3.2. For example, if the CPC

"Working conditions" is rated as "Advantageous", it has a "Positive" effect on the situation; all CPCs are evaluated in the same way, and the number of CPCs whose effect is found to be "Negative" or "Positive" is obtained in the analysis.

Table 3.2: Specification of CPCs.

CPCs	Level	Effect	CPC	Level	Effect
1. Availability of resource	Adequate	Positive	2. Adequacy of Training and experience	Adequate	Positive
	Adequate with limitations	Not Significant		Adequate with limitations	Not Significant
	Inadequate	Negative		Inadequate	Not Significant
3. Quality of communication	Very efficient	Positive	4. HMI (Human Machine Interaction)	Supportive	Positive
	Efficient	Not Significant		Adequate	Not Significant
	Inadequate	Negative		Tolerable	Not Significant
	Deficient	Negative		Inappropriate	Negative
5. Access to procedures	Appropriate	Positive	6. Working conditions	Advantageous	Positive
	Acceptable	Not Significant		Compatible	Not Significant
	Inappropriate	Negative		Incompatible	Negative
7. Number of simultaneous goals	More than Capacity	Positive	8. Available time	Adequate	Positive
	Matching Capacity	Not Significant		Temporarily Inadequate	Not Significant
	Less than Capacity	Negative		Inadequate	Negative
9. Circadian rhythm	High	Positive	10. Crew collaboration quality	Very Efficient	Positive
	Moderate	Not Significant		Efficient	Not Significant
	Low	Negative		Inefficient	Not Significant
				Deficient	Negative
11. Adequacy of organization	Very Efficient	Positive			
	Efficient	Not Significant			
	Inefficient	Not Significant			
	Deficient	Negative			

Some CPCs are interrelated with each other, and their relationships are shown in Table 3.3. According to these relationships, the effect of CPCs is updated according to dependencies among CPCs after a set of all the CPCs' effects is identified. For instance, if an effect of CPCs in the left column of Table 3.3 is "Not Significant", and more than three or four CPCs in the right column are "Positive" or "Negative", the effect of CPCs in the left column also becomes "Positive" or "Negative", respectively.

Table 3.3: Dependencies among CPCs.

CPC	Affected by the following CPCs
Working conditions	Adequacy of organization, Adequacy of HMI, Circadian rhythm, Available time, Adequacy of training and experience
Number of simultaneous goals	Working conditions, Adequacy of HMI, Access to procedures
Available time	Working conditions, Adequacy of HMI, Access to procedures, Number of simultaneous goals, Adequacy of training and experience, Circadian rhythm
Crew collaboration quality	Adequacy of organization, Circadian rhythm, Quality of communication

On the basis of the set of CPC effects and the chart shown in Fig. 3.7, the control mode is identified. The control mode originates from Contextual Control Model (COCOM) (Hollnagel, 2003). In COCOM, the control is described in terms of how we do things and how actions are chosen and executed according to their surrounding environment. Moreover, the outcome of the control prescribes a certain sequence of the possible actions, and the sequence is constructed by the environment rather than pre-defined. The control can obviously occur on several levels or in several modes, and in this context, four linguistic values shown in Table 3.4 have been introduced as the control mode. representing how actions of human beings are chosen and executed according to their surrounding environment Also, each control mode is related to an interval of probability of action failure (PAF), as shown in Table 3.5. Here, note that the chart in Fig. 3.7 carries the premise that the weights of CPCs, which represent the significance of those CPCs for a subject of CREAM analysis, are all equivalent.

Table 3.4: Description of control mode (Hollnagel, 1998, 2003).

Control Mode	Each control mode represents situations in which:
Strategic	Tasks can be carried out based on sophisticated strategies and enough comprehension of global context.
Tactical	People can carry out their tasks based on planning and following procedures, but their scope is limited.
Opportunistic	The next action heavily relies on the salient features of a current context rather than stable strategies.
Scrambled	People are too upset to choose a proper action; an extreme case here would be the state of momentary panic.

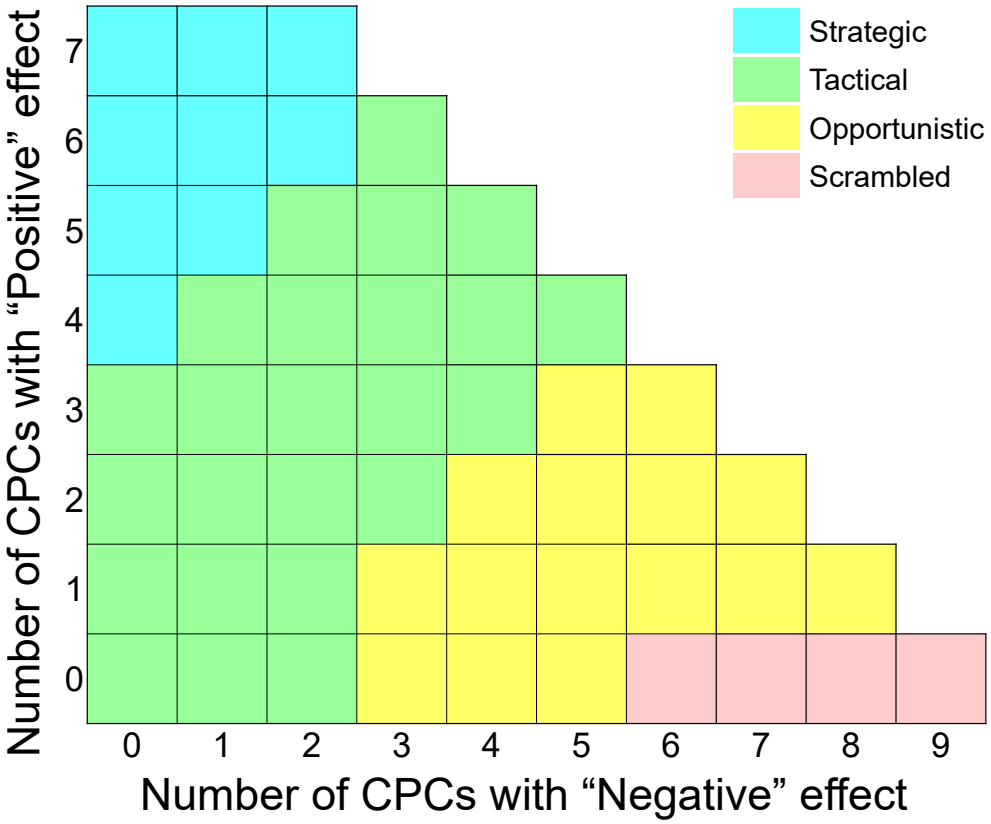


Fig. 3.7: Relation between CPC effect and control modes.

Table 3.5: PAF intervals with respect to control modes.

Control Mode	Intervals of probability of action failures
Strategic	$0.50 \times 10^{-5} < p < 0.010$
Tactical	$0.10 \times 10^{-2} < p < 0.10$
Opportunistic	$0.010 < p < 0.50$
Scrambled	$0.10 < p < 1.00$

For further development of the CREAM, several studies have introduced fuzzy logic theory into the original CREAM to make it continuous and quantitative; these are generally called Fuzzy CREAM. In Fuzzy CREAM, membership functions of CPC levels whose support set is CPC score are defined. The CPC score is a continuous value varying from 0 to 100 that represents the status of the CPC: the higher the CPC score, the better the CPC status. These membership functions also represent the degree of matching between a specific CPC score and a particular CPC level, varying from 0 to 1.00. Four membership functions of the control modes, whose support set is the logarithm of PAF, are defined in the same way. Then, fuzzy linguistic rules,

which are $IF-THEN$ rules between combinations of CPC levels and a specific control mode, are defined, e.g.,

$$IF\, S_1 = Compatible\, AND\, S_2 = Efficient\, \cdots\, AND\, S_m = \cdots\, THEN\, C = Strategic,$$

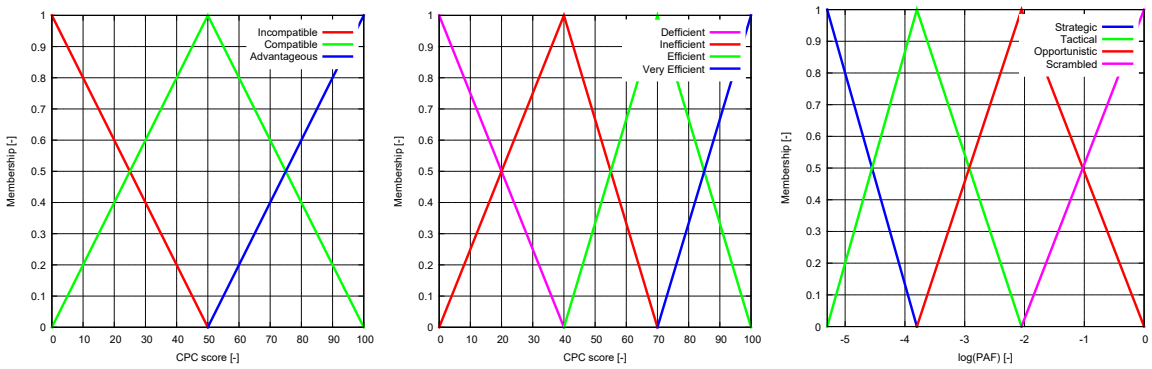
where S_i denotes the level of the i-th CPC, m is the total number of CPCs, and C represents the control mode ($1 \leq i \leq m$). With the above items, a conclusion fuzzy set of the control mode is consequently obtained by calculating how the antecedent matches the consequent in those $IF-THEN$ rules.

There exist several Fuzzy CREAM models based on this fundamental idea. Konstandinidou et al. (2006), for example, constructed 46,656 fuzzy rules by hand using the chart in Fig. 3.7 and calculated a conclusion fuzzy set with the min-max inference technique (George et al., 1995). Also, Yang et al. (2013) defined the relative weight of CPCs and calculated the belief degrees of each control mode with the Bayesian network of CPCs.

In the proposed FRAM model, the *weighted CREAM model* (Ung, 2015), which is also one of the Fuzzy CREAM models, is adopted, for two reasons. First, our method considers the concept of CPC weight; while the weight must be different case by case, it is regarded as equal in many cases for the sake of simplicity. Second, the table in Fig. 3.7 is not necessary, as the chart in Fig. 3.7 would not be available if the CPC weights were not equivalent. A model that satisfies both of these requirements is quite rare, and the *weighted CREAM model* is one of them. The algorithm consists of the following four steps:

Step 1: Definition of Membership Functions for Linguistic Values of CPC Levels

The first step of this model is to define membership functions. Examples are shown in Fig. 3.8. Figure 3.8(a) and 3.8(b) show examples of the membership functions of CPC levels. Figure 3.8(c) shows the membership functions of control modes, along with the logarithm of the probability, whose base is 10, which is used for their support set because the lower limit of PAF is assumed to be 0.50×10^{-5} according to the original CREAM (Table 3.5). Ideally, the membership functions should be designed with statistical data and/or with the knowledge of experts. However, in this paper they are all regarded as the simple triangular functions shown in Fig. 3.8 for the sake of simplicity.



(a) Working conditions (b) Quality of communication (c) Control modes

Fig. 3.8: Examples of membership functions.

Step 2: Construction of Fuzzy Rules

Fuzzy rules are constructed with a systematized process in this step. Ideally, a rule should be obtained with the statistical data and/or the knowledge of experts. However, since each CPC has three or four CPC levels, as shown in the example in Fig. 3.8, tens of thousands of combinations of CPC levels are obtained as antecedents, i.e., the IF-part of the rule. For example, two levels are identified with respect to a CPC score in the case of Fig. 3.8, generating 2^{11} combinations of CPC levels as a result of evaluating all nine CPC scores. Therefore, systematic ways to distribute those combinations to a specific control mode, $THEN$-part, and obtain an $IF-THEN$ rule as shown below are required.

$$IF \begin{bmatrix} S_{1,1} = \cdots & AND & \cdots & AND & S_{1,m} = \cdots \\ S_{2,1} = \cdots & AND & \cdots & AND & S_{2,m} = \cdots \\ \vdots & & \ddots & & \vdots \\ S_{l,1} = \cdots & AND & \cdots & AND & S_{l,m} = \cdots \\ \vdots & & \ddots & & \vdots \\ S_{n,1} = \cdots & AND & \cdots & AND & S_{n,m} = \cdots \end{bmatrix} THEN\ C = C_k$$

Here, n is the total number of combinations of CPC levels belonging to the k-th control mode: C_k $(1 \leq k \leq 4)$. Also, $S_{l,i}$ represents a level of the i-th CPC in the l-th combination of CPC levels in the IF-part $(1 \leq l \leq n)$.

An index I^l is introduced $(0 \leq I^l \leq 100)$ to distribute a combination of CPC levels to a specific control mode. The index is defined as

$$I^l = \sum_{i=1}^{m} A_i^l \cdot w_i, \tag{3.1}$$

where A_i^l is the significance of the i-th CPC level in the l-th combination of the IF-part ($0 \leq A_i^l \leq 100$). It is defined as a value on the abscissa where the membership function reaches 1.00. For example, the significance of "Advantageous", "Compatible", and "Incompatible" is 0, 50, and 100, respectively in Fig. 3.8 (a). Also, w_i is the normalized relative weight of the i-th CPC ($0 \leq w_i \leq 1.00$) with the following equation:

$$w_i = \frac{W_i}{\sum_{i=1}^{m} W_i}, \tag{3.2}$$

where W_i is the relative weight of the i-th CPC set by analysts ($W_i \geq 0$). I^l is regarded as a percentage on the abscissa in Fig. 3.8(c); the value on a specific point of the abscissa, SV^l, is identified by following equation ($-5.30 \leq SV^l \leq 0$).

$$SV^l = -5.30 \times \frac{I^l}{100}. \tag{3.3}$$

A combination of CPC levels, i.e., the IF-part, belongs to a specific control mode C_k, i.e., the $THEN$-part, depending on the intervals listed in Table 3.6 to which the SV^l belongs.

Table 3.6: Relationships between intervals of $log(PAF)$ and control mode.

Control Mode	Strategic	Tactical	Opportunistic	Scrambled
Interval of $log(PAF)$	[-5.30, -3.80]	(-3.80, -2.90]	(-2.90, -1.03]	(-1.03, 0]

Step 3: Acquisition of Fuzzy Conclusion

In this step, a conclusion fuzzy set of the control mode is obtained by the calculation of μ^{C_k}, the degree of matching for each control mode, which is obtained by

$$\mu_l^{C_k} = \sum_{i=1}^{m} \mu_{l,i}^{C_k}(x) \cdot w_i, \tag{3.4}$$

$$\mu^{C_k} = \frac{\sum_{l=1}^{n} \mu_l^{C_k}}{n}, \tag{3.5}$$

where $\mu_{l,i}^{C_k}(x)$ is the value of the membership function corresponding to the level of the i-th CPC in the l-th $THEN$-part whose linguistic consequent is C_k, all of which vary from 0 to 1.00. Also, x is the CPC score ranging from 0 to 100.

The concluding fuzzy set $\mu(y)$ is obtained by using μ^{C_k}, which is defined as

$$\mu(y) = min(max(\nu^{C_1}(y), \mu^{C_1}), max(\nu^{C_2}(y), \mu^{C_2}), \cdots, max(\nu^{C_k}(y), \mu^{C_k}), \cdots), \tag{3.6}$$

where $\nu^{C_k}(y)$ is the membership function of the k-th control mode shown in Fig. 3.8(c) ($0 \leq \nu^{C_k}(y) \leq 1.00$), and y equals $log(PAF)$ varying from -5.30 to 0.

Step 4: Defuzzification

The concluding fuzzy set is transformed into a crisp value by the following defuzzification process:

$$CV = \frac{\int_D y \cdot \mu(y) dy}{\int_D \mu(y) dy}, \tag{3.7}$$

where CV is the crisp value of $log(PAF)$ and D is the domain of integration. In other words, the crisp value corresponds to the center of gravity of the fuzzy set.

Based on the idea of Fuzzy CREAM, the proposed FRAM model assumes that a state in each function can be described in terms of the control mode, and the state can change according to the state of their surrounding environment created by CPCs and their weight; although Hollnagel (2012a) pointed that the CREAM is obsolete, it is still useful to represent the ETTO principle. The variabilities in each function and their surrounding environment are eventually defined as the change of their continuous control mode (crisp value of PAF) and CPC scores, respectively.

These definitions are based on the fact that the origin of the FRAM functions is derived from the framework of Work Domain Analysis (WDA) (Rasmussen et al., 1994; Vicente, 1999). In this framework, the functions are described as predicates and essentially qualitative properties. The functional safety in this context refers to the safety of those functions and is therefore described as qualitative properties as well. That is why the control mode can be introduced to represent such qualitative properties, and moreover, the representation, i.e., control mode is now supported by semi-quantitative criteria based on Fuzzy CREAM as well; those definitions of variabilities consequently therefore enable us to investigate the functional safety on the basis of the balance between qualitative and quantitative perspectives.

3.2.2 Formulating Interaction among Functions and Surrounding Working Environment

The definition of variabilities enables us to formulate interactions among functions. First of all, the propagation of a functional variability in an upstream function propagates to its downstream functions as follows:

$$x_{i,down}^{t+1} = \frac{PAF_{up}^{t}}{PAF_{up}^{t+1}} \times x_{i,down}^{t}, \tag{3.8}$$

where $x_{i,down}^{t+1}$ and $x_{i,down}^{t}$ are the updated scores and original scores of the i-th CPC in a downstream function, respectively. Also, PAF_{up}^{t} and PAF_{up}^{t+1} referring to the PAF

value of a particular upstream function before and after the PAF value, respectively, have been changed by the Fuzzy CREAM process. Equation 3.8 represents that some specific CPC scores in the downstream functions decrease if the PAF in an upstream function increases, and vice versa.

The definition of variabilities also makes it possible to formulate dependencies among the CPCs in Table 3.3, as follows:

$$x_i^{t^*+1} = x_i^{t^*} + \sum_j (x_j^{t^*} - x_i^{t^*}) \times w_j, \tag{3.9}$$

where $x_i^{t^*+1}$ is an updated score of the i-th CPC in the left column of Table 3.3, and $x_i^{t^*}$ is its original score. Also, $x_j^{t^*}$ is the score of the j-th CPCs listed in the right column of Table 3.3, and w_j is the normalized weight of the CPC. This calculation process is hereinafter called CPC resonance.

The unitary mechanism of one function is implemented as shown in Fig. 3.9 on the basis of the above definitions and equations. That is, manual interventions or variabilities coming from upstream functions change the original CPC scores as a trigger first. The scores are then updated according to the dependency among CPCs by Eq. 3.9, which is the input of the Fuzzy CREAM process. The Fuzzy CREAM process then update PAF in the function based on the input CPC scores. The induced functional variability, i.e., change of PAF propagates to downstream functions on the basis of Eq. 3.8, changing the CPC scores of downstream functions.

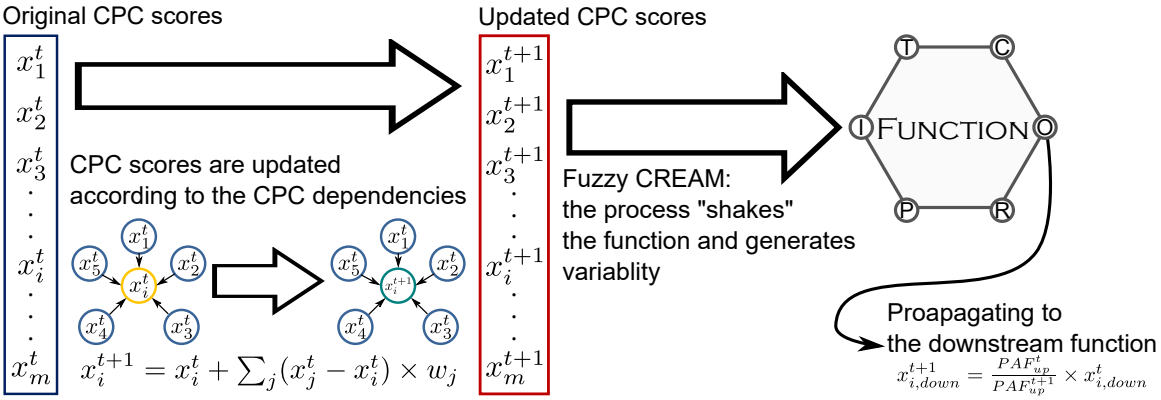


Fig. 3.9: Unitary mechanism of one function.

More schematic representation of the mechanism is shown in Fig. 3.10. According to Fig. 3.10, a set of CPCs, i.e., context surrounds a function to induce its variability while it is, at the same time, affected by variabilities propagating from upstream functions through aspects of the function. Here, it should be noted that the CPC

scores being affected by variabilities coming from upstream functions are determined both by the aspects of a downstream function and by characteristics of their upstream functions. These unitary mechanisms are connected as shown in Fig. 3.11 in the end.

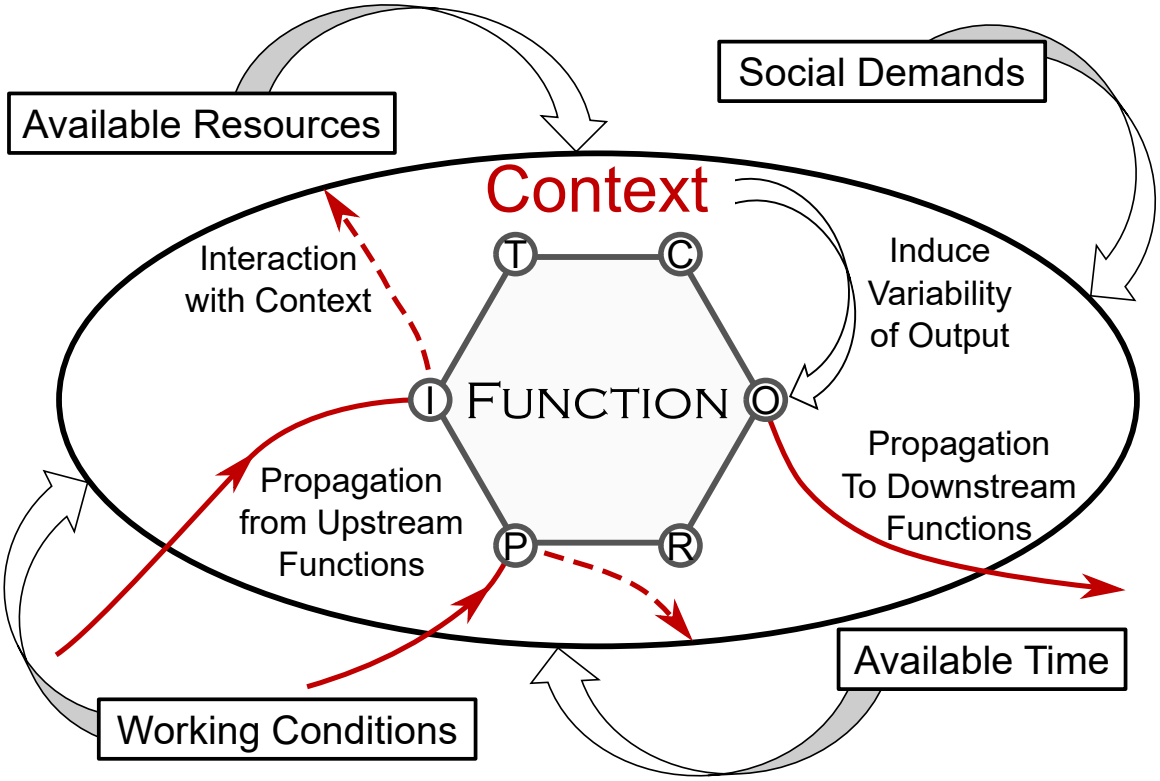


Fig. 3.10: Schematic illustration of unitary mechanism of function.

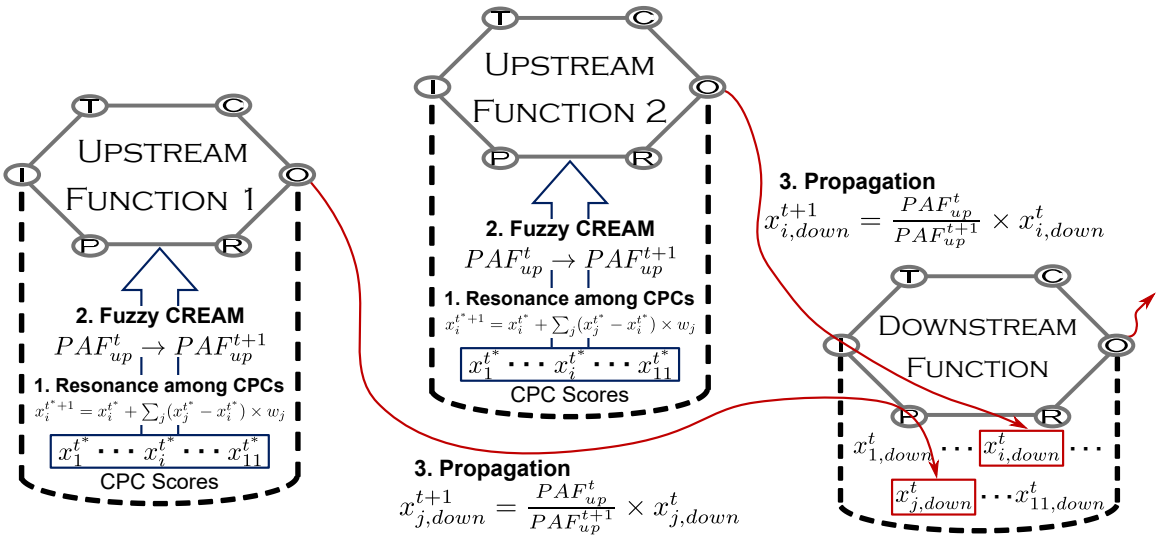


Fig. 3.11: Overview of initial model: connecting units shown in Fig. 3.9.

The flowchart of this model is shown in Fig. 3.12. After the initial setting is finished, potential couplings between functions are instantiated. Then, once some CPC scores in the functions are changed manually as a trigger of the simulation, the process shown in Fig. 3.11 is repeated recursively. This process provides snapshots of how the variabilities of working environment influence the functional variabilities, and their propagation changes the performance of other functions. After the process is finished, the obtained result can be the initial state of next simulation. That is, it is possible to change the CPC scores of the result and simulate again, enabling the interactive analysis. Note that the process continues automatically after the CPC scores are changed manually at the beginning of the simulation process. In other words, what should be done manually are initial setting and changing the CPC scores as a trigger of the process, and there is no room for intervention after that. This initial model is applied to the investigation of an actual aircrash accident, whose detail is presented in the chapter 4.

In the end, it is limited for the current version of the FRAM model to describe the evolutions of socio-technical systems since this model just provides snapshots of them. However, the resilience engineering requires the FRAM to investigate dynamic/continuous characteristics of socio-technical systems rather than static/discrete pictures of the systems. This model has been updated in response to this demand, and the detail is shown in the next section.

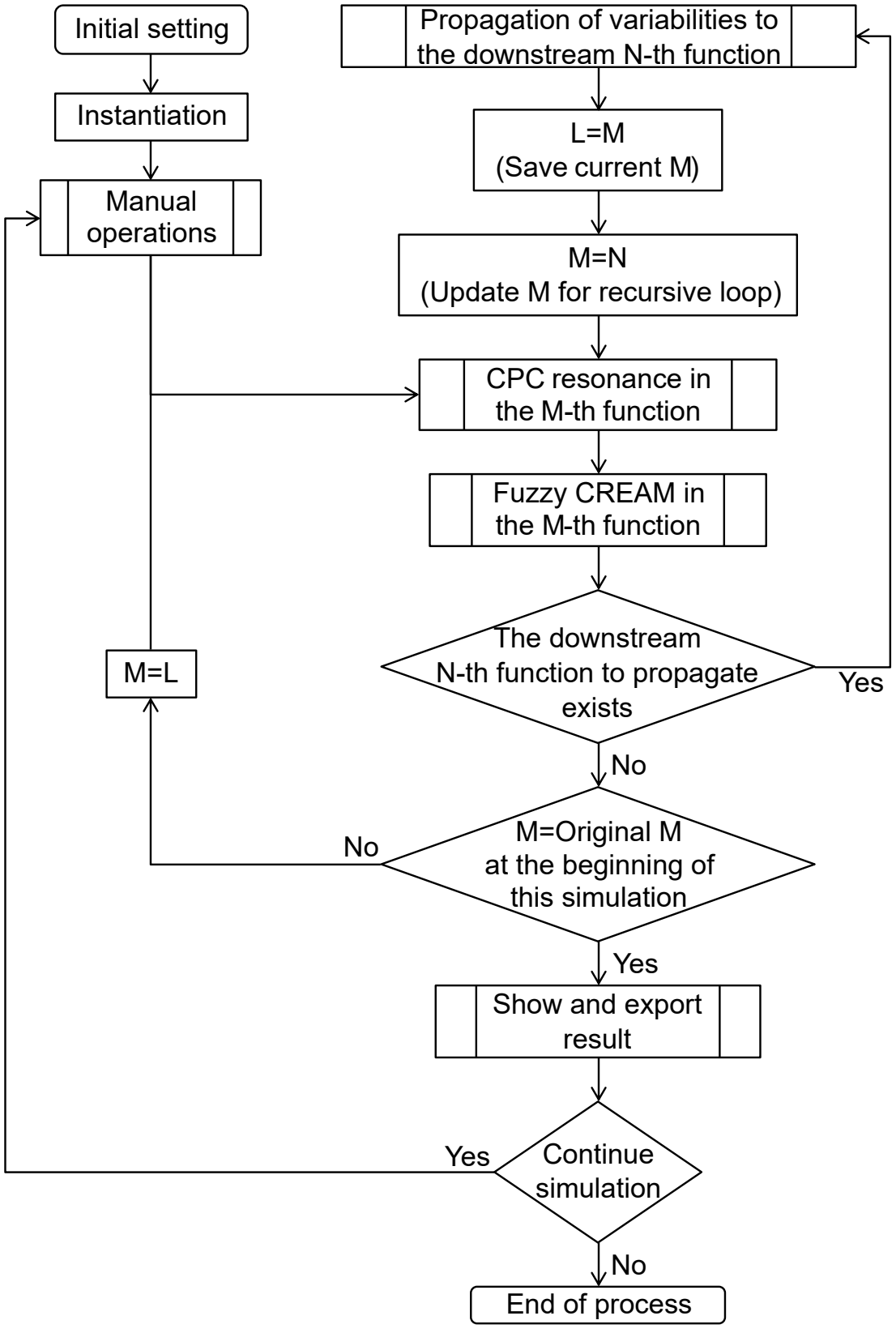


Fig. 3.12: Flowchart of initial FRAM model.

3.3 Extended FRAM Model Based on Cellular Automaton

The initial FRAM model shown in the previous section has been extended to overcome its limitations. This update is to modify the algorithm shown in Fig. 3.12 in accordance with the concept of Cellular Automaton (Neumann et al., 1966); the main structure of functions shown in Fig. 3.9 – 3.11 remains the same. This modification enables to reveal the dynamical transition of safety in each function.

Cellular Automaton has initially been introduced by Neumann et al. (1966). In trying to represent data processing in a machine that mimicked the self-reproducing that goes on in life, he discovered a kind of simulation carried out on a grid. In this process, the state of each cell on the grid is supposed to change every moment depending on the states of surrounding neighborhoods. Moreover, simple local rules to update each cell interact with each other and keep on changing the state of those cells. This is known as automaton on the grid in parallel automatically. What is interesting with this simulation is that the interaction of simple local rules could bring about a very complex macroscopic behavior: Cellular Automaton can execute basic logical operations AND; OR; NOT and work as if it is a computer. However, the original model that John von Neumann came up with was too complicated for practical use since each of the cells is supposed to have 29 states on the grid. More simplified models have since been built.

Conway's Game of Life (Berlekamp et al., 1982) is one of the best-known models of Cellular Automaton. In this model, every cell is supposed to have only two states, dead or alive, and they change their state simultaneously in parallel according to the following simple local rules (Fig. 3.13).

Rule 1: An alive cell dies in the next generation if it is surrounded by less than two alive neighbor cells, as if it were caused by underpopulation.

Rule 2: An alive cell can survive in the next generation if it is surrounded by two or three alive neighbor cells.

Rule 3: An alive cell dies in the next generation if it is surrounded by more than three alive neighbor cells, as if it were caused by overpopulation.

Rule 4: A dead cell can revive if it is surrounded by exactly three alive neighbor cells, as if it were a reproduction.

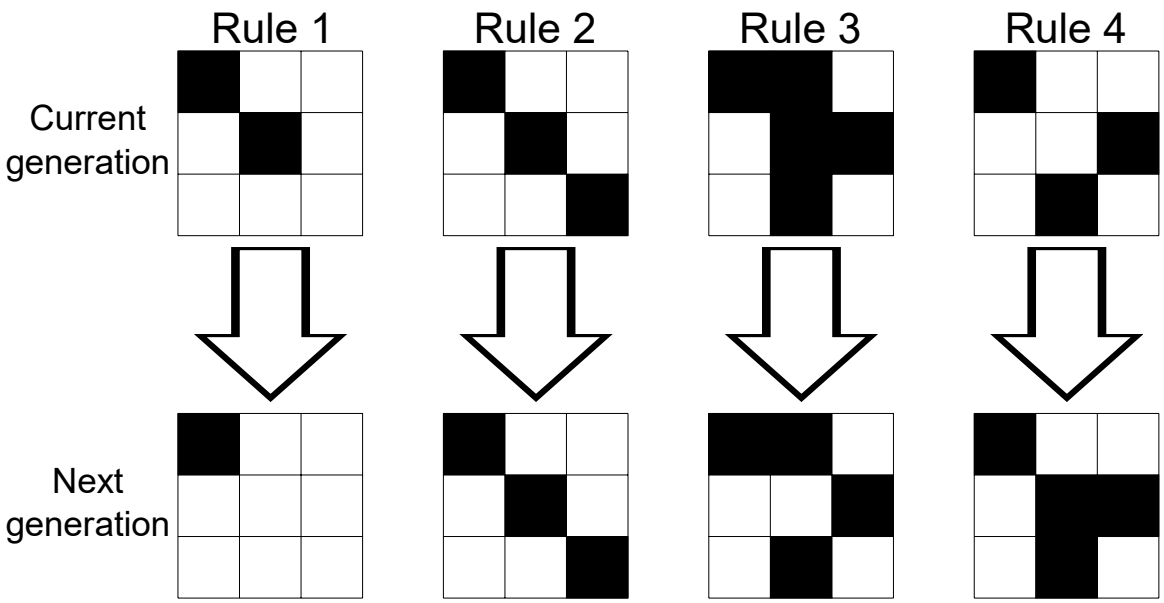


Fig. 3.13: Rules of Conway's Game of Life.

The local transition of each cell creates unusual global behavior; a number of characteristic patterns can be observed in the process, including complex behavior, as if it is a computer.

These simulations are now applied to various fields to investigate complex phenomena. One of the most famous examples is the simulation of traffic flow to investigate how dynamic traffic behavior, such as traffic jams, will go on (Fukui and Ishibashi, 1996; Qian et al., 2017; Ruan et al., 2017). In other cases, they are also applied to simulations of pedestrian group behavior during emergency evacuation (Fu et al., 2015; Lu et al., 2017). Moreover, the Cellular Automaton can also be applied to the field of materials science to investigate the dynamic behavior of materials such as deformation (Liu et al., 2017). The simulation of Cellular Automaton is indeed applicable to a wide variety of fields, and it is also the case with the resilience engineering.

This mechanism of Cellular Automaton has been introduced into the updated FRAM model. In this model, FRAM functions are regarded as cells, and the calculation processes shown in Fig. 3.9 or 3.10 correspond to the simple local rules of Cellular Automaton. Moreover, every function in a FRAM network runs and repeats those processes simultaneously in parallel; thus, a dynamic changing pattern of the PAFs of the functions is obtained. This is the significant difference compared to the previous model, which had no consideration of time transitions or parallel processing of functions. The extended FRAM model is expected to represent complex behaviors similar to Cellular Automaton.

The entire flow of this model is shown in Fig. 3.14. The process starts with setting the initial parameters, including functions, their aspects, their dependencies, instantiation, and so on; the details will be shown in the next section with an actual example. Then, the process moves on to the manual operations, where CPC scores in functions or dependency among functions can be modified. The manual operation of setting the initial values of CPS scores for the functions triggers the main process during which states of functions change on the basis of the mechanism shown in Fig. 3.11, similar to Conway's Game of Life. In other words, three main processes — CPC resonance, Fuzzy CREAM, and propagation of variabilities among functions — are carried out in this order in one generation. This process is autonomously iterated until the calculation process is cut in, and the process go back to the manual operations again with the intervention.

Here, it should be noted that the dimensionless simulation time: $T[-]$ in Fig. 3.14 is different from the real time; it is defined as the number of loops in which a set of the three main processes is repeated. The concept of simulation time: $T[-]$ is rather consistent with that adopted in qualitative reasoning (Kuipers, 1986, 1994, 2001), in which the time-points are defined as those points in time when a qualitative state of the simulation model (i.e., the qualitative value of any variables) changes. Therefore, the time in the simulation world goes on independent of that in the real world; the very long time in the simulation world can corresponds to a moment in the real world, and vice versa. The simulation time: $T[-]$ hereafter advances based on this idea; the qualitative change is brought about by the loops of the above three main processes, and the loops advance the dimensionless simulation time: $T[-]$.

This model is applied to an investigation of supply chain operations, focusing on their complexity and dynamic behavior. The detail is shown in the chapter 5.

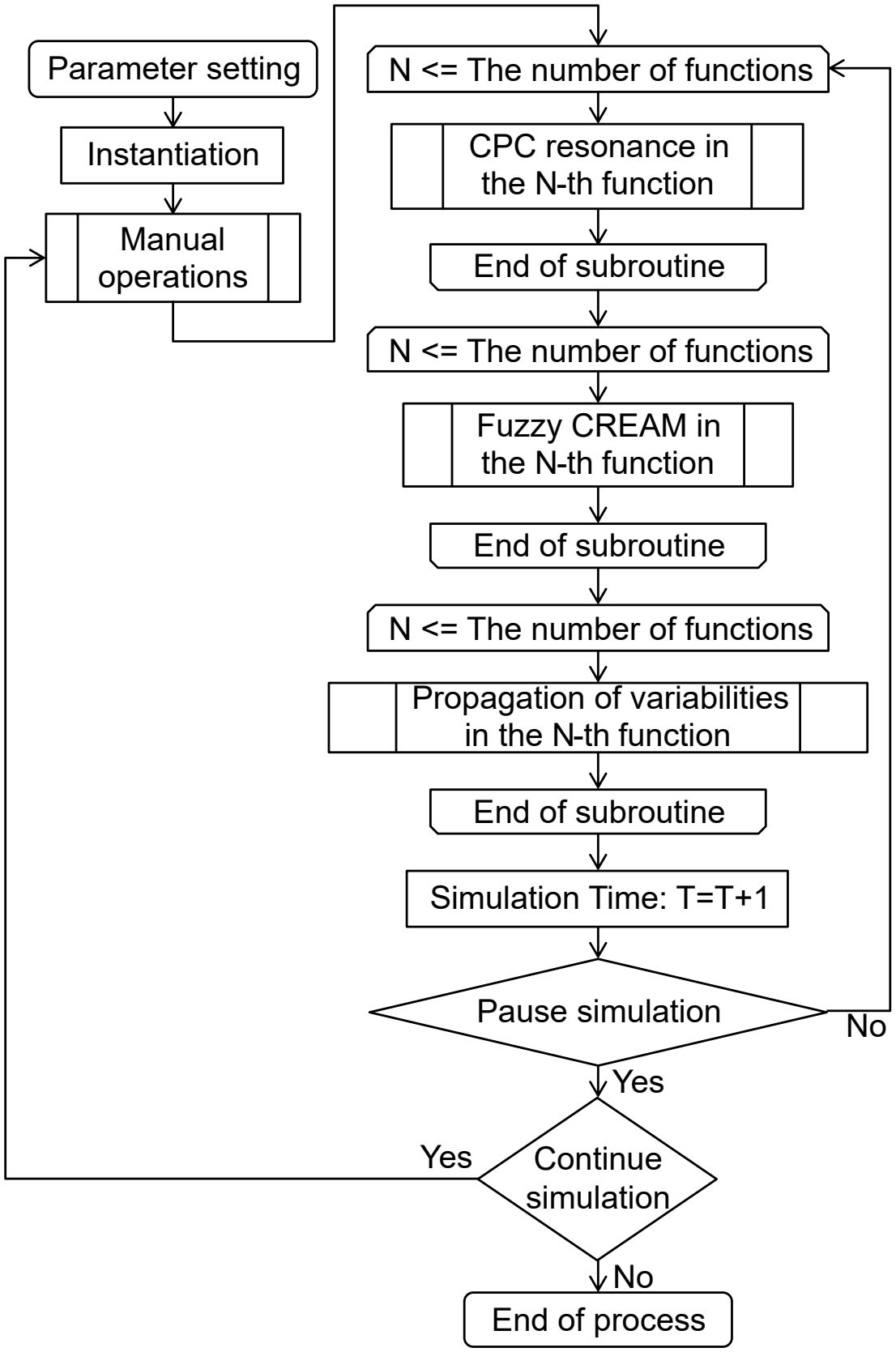


Fig. 3.14: Flowchart of FRAM model based on Cellular Automaton.

3.4 Extended FRAM Model Based on Structure of Complex Adaptive Systems

3.4.1 Inconsistency of FRAM Model Structure

The previous two FRAM models have been developed so that a context consisting of eleven CPCs surrounds one function, and they interact with each other, according to Fig. 3.10. However, these models put too much focuses on the relationship between a context and single function. According to the resilience engineering and FRAM, the context is an entity which globally interacts with all functions as shown in Fig. 3.15, rather than a local parameter specific to a single function. That is why the FRAM model has been further updated so that the model structure is consistent with the concept.

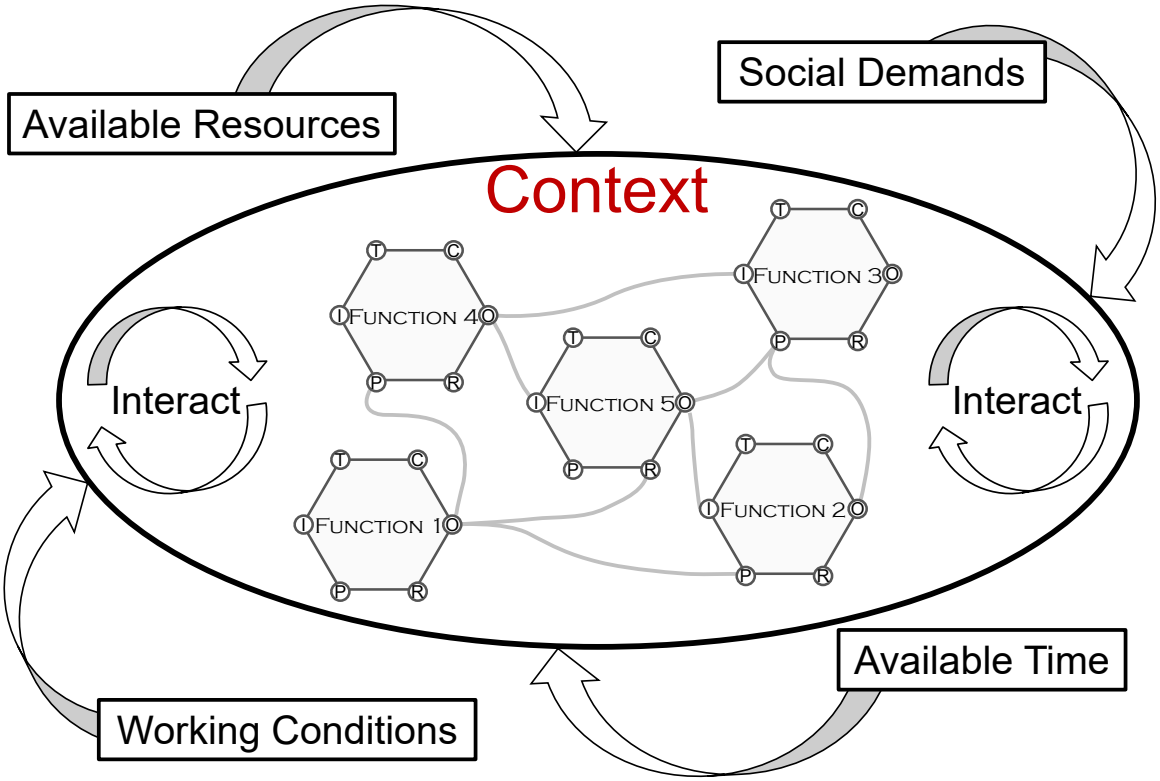


Fig. 3.15: Basic concept to update FRAM model.

3.4.2 Revision of Previous FRAM Model

Figure 3.16 illustrates an overview of the updated FRAM model. The crucial difference from the previous model in Fig. 3.11 is its structure where a set of CPC scores

is shared by each FRAM function of an instance, and the set of CPC scores simultaneously induces the variability of FRAM functions. The algorithm of this model is almost the same as the previous — the cellular automaton based model. That is, the numerical unbalance of CPC scores updates themselves based on the CPC dependency, and those scores updates PAF of functions. Also, the updated PAF in a function propagates to downstream functions and interacts with the shared set of CPC scores again.

It should be noted that Table 3.6, which is necessary for Fuzzy CREAM, has also been updated as shown in Table 3.7 on this occasion. This is because Table 3.6 involves some arbitrariness of the original research (Ung, 2015). In contrast, the updated table was obtained more systematically by applying OR operation to membership functions shown in Fig. 3.8(c).

Table 3.7: Updated relationships between intervals of $log(PAF)$ and control mode.

Control Mode	Strategic	Tactical	Opportunistic	Scrambled
Interval of $log(PAF)$	[-5.30, -4.55]	(-4.55, -2.93]	(-2.93, -1.03]	(-1.03, 0]

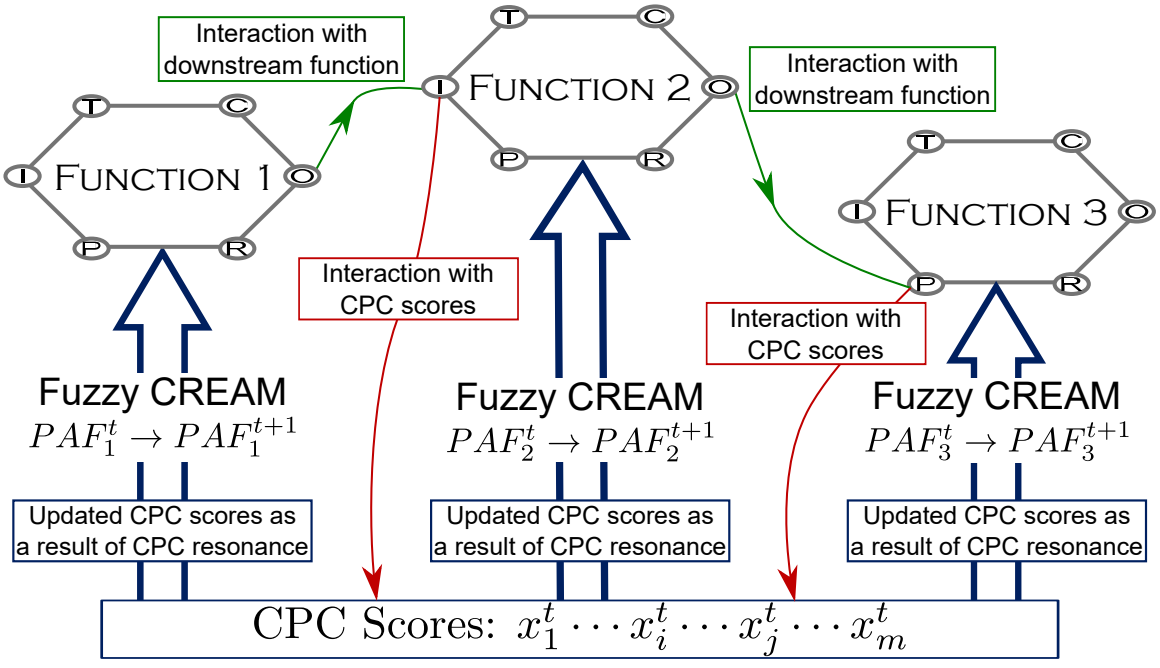


Fig. 3.16: Overview of revised FRAM model based on Fig. 3.15.

In this model, the CPC resonance is formulated as the following equation:

$$x_i^{t+1} = x_i^t + \frac{\sum_{j=1}^{k}(x_j^t - x_i^t) \times w_j}{\sum_{j=1}^{k} w_j}, \tag{3.10}$$

where x_i^{t+1} is an updated score of the i-th CPC in the left column of Table 3.3, and x_i^t is its original score; x_j^t and w_j are the score and weight of the j-th CPCs in the right column of Table 3.3, respectively; k is the number of surrounding CPCs listed in the right column of Table 3.3. The Eq. 3.9 calculates the weighted average of those CPCs, and a set of the updated scores will be the input value of the Fuzzy CREAM as well as other two models.

In addition, the propagation of functional variabilities to downstream functions, represented as the green lines in Fig. 3.16, is formulated as following:

$$PAF_{down}^{t+1} = PAF_{down}^{t} + \frac{\sum_{j=1}^{l}(PAF_{up,j}^{t+1} - PAF_{up,j}^{t})}{l}, \tag{3.11}$$

where PAF_{up}^{t} and PAF_{up}^{t+1} respectively refer to the crisp PAFs in a upstream function before and after the PAF is updated by the Fuzzy CREAM process; this is also the case with PAF_{down}^{t} and PAF_{down}^{t+1}. Also, l is the number of functions surrounding a certain downstream function. In this calculation, the difference of PAFs between a subject downstream function and upstream functions is accumulated and averaged with the number of the upstream functions; the result is added to the original PAF in the downstream functional in the end.

The effect of interactions among the functions is then looped back to the set of CPC scores, which is represented as the red lines in Fig. 3.16. In other words, the fluctuation of CPC scores changes crisp values of PAF in each function, and the CPC scores themselves are also affected by its effect again. This is formulated as following:

$$x_i^{t+1} = \frac{PAF_{down}^{t}}{PAF_{down}^{t+1}} \times x_i^{t} \tag{3.12}$$

where x_i^{t+1} and x_i^t are the updated and original scores of the i-th CPC, respectively. Here, the affected CPCs depend on the aspects where the propagating variabilities go through, and the relationships are defined as shown in Table 3.8.

Table 3.8: Dependency among the new set of CPCs.

Aspects that variabilities from upstream functions propagate through	ID No. of CPCs in Table 3.2										
	1	2	3	4	5	6	7	8	9	10	11
Input	1	1	1	0	1	1	1	1	0	1	0
Precondition	1	1	1	0	1	1	1	1	0	1	0
Resource	1	0	0	0	0	0	0	1	0	0	0
Control	1	1	1	0	1	1	1	1	0	1	0
Time	0	0	0	0	0	0	0	1	0	0	0

The entire flow of this model is summarized as a flowchart shown in Fig. 3.17. The initial setting includes the parameter setting and instantiation process, whose detail will be shown in the next section. The next step is the manual operation including artificial manipulation of CPC scores or changing dependency among functions, i.e., further instantiations. Once these steps are completed, the process of CPC resonance formulated as Eq. 3.10, Fuzzy CREAM, and interactions of variabilities formulated as Eq. 3.11 and Eq. 3.12 are repeated; it should be noted that the practitioners basically cannot interfere the repetitive process and its result, and it goes on automatically until they cut in by hitting a key on the keyboards. The process goes back to the manual operation if the practitioners cut in the repetitive process, and this enables them to examine the effect of additional variabilities to the result of the previous simulation. Based on these, a sequential change of the crisp PAFs in each function is obtained, and it provides generic insights about the safety of socio-technical systems.

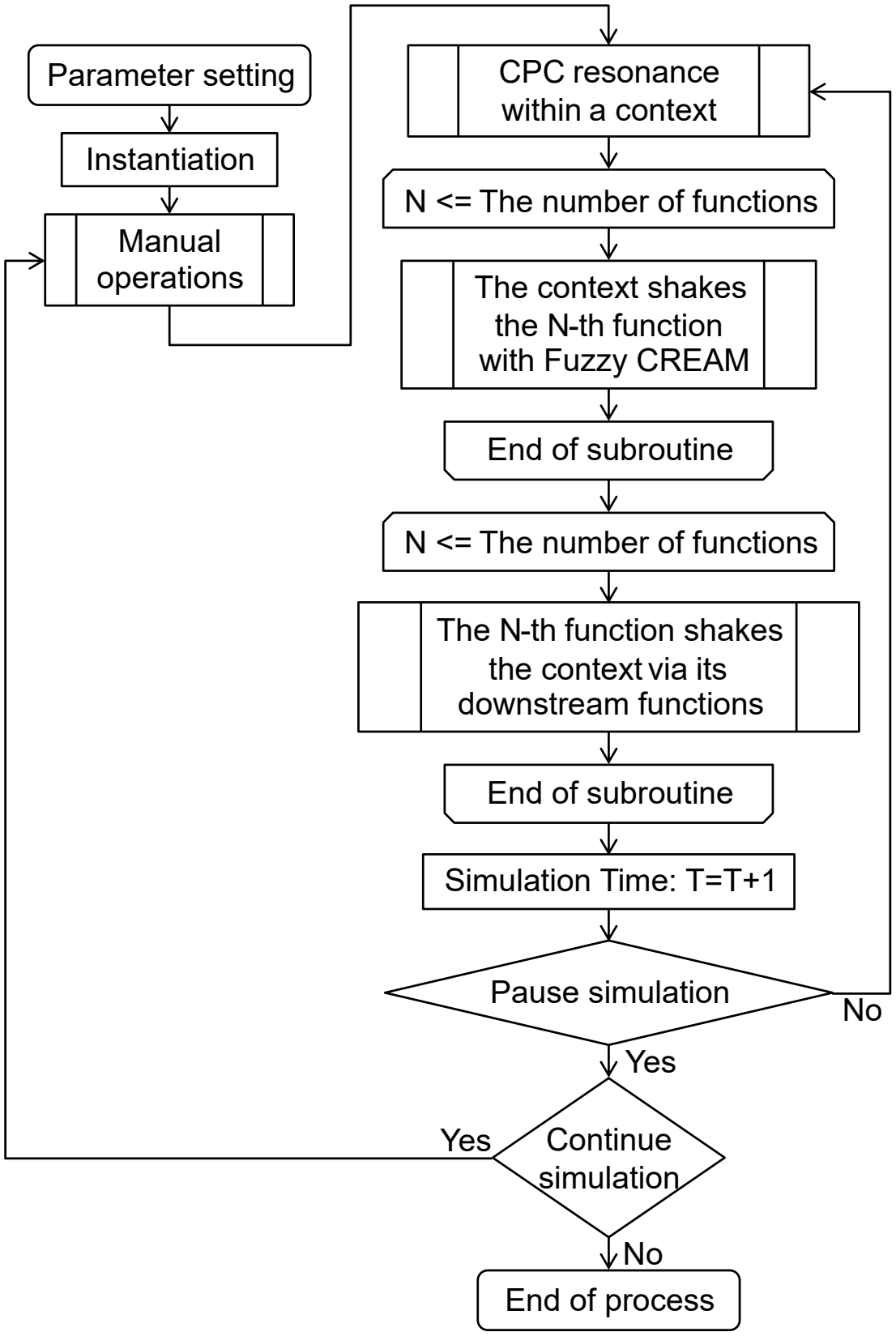


Fig. 3.17: Flowchart of revised FRAM model shown in Fig. 3.16.

3.4.3 Another Perspective: FRAM Model Based on Structure of Complex Adaptive Systems

The updated model can be described from another perspective. The model structure shown in Fig. 3.16 suggests that a specific working environment or context can be regarded as a space spanned by a vector of CPC scores, and an instance are included in the context as shown in Fig. 3.18. This model, therefore, has a structure in which the context and instance mutually "shakes" in response to the other, and their mutual interaction results in dynamical change of the safety of each function. This is typical dynamics specific to Complex Adaptive Systems (Mitchell, 2009; Johnson, 2009), and the model now obtained a structure consistent with the original idea of the FRAM and resilience engineering.

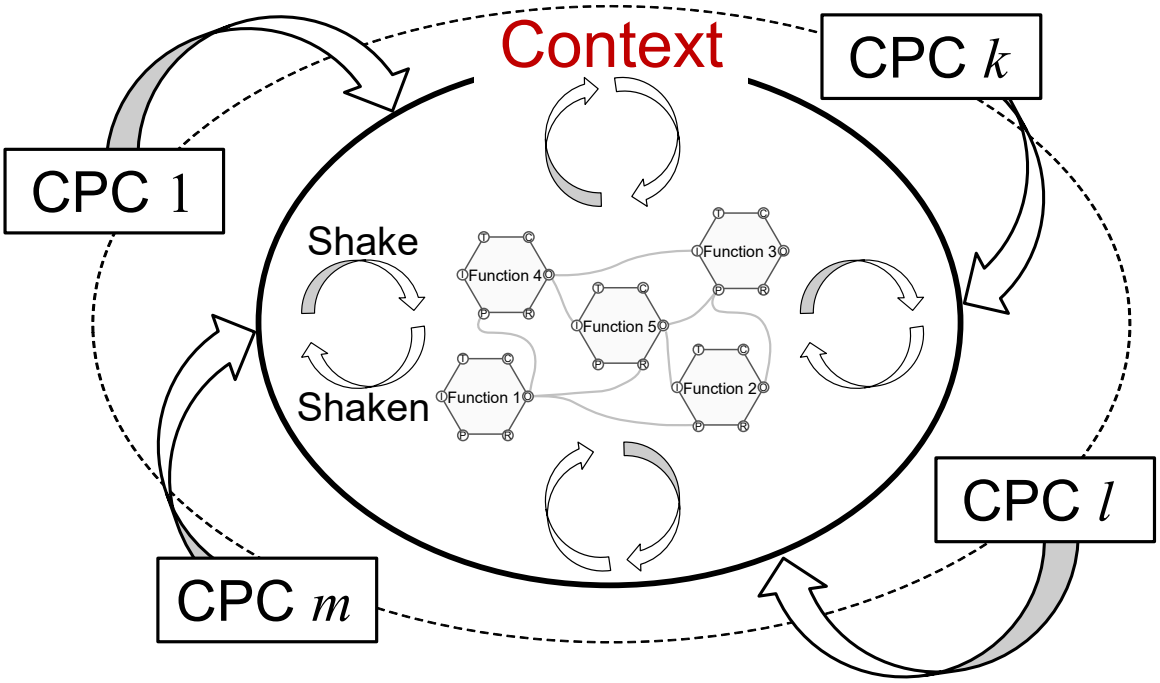


Fig. 3.18: Another perspective of the FRAM model shown in Fig. 3.16.

This model is applied to a safety analysis of conditional driving automation issued by SAE (2016). This case study is to investigate the safety and feasibility of a new coming technology which is about to be developed and implemented on the society. In other words, the investigation addresses the most challenging aspect of the envisioned world problem. The detail is provided in the chapter 6.

Chapter 4

Validity of Compensatory Principle in Ever Changing Environment

4.1 Introduction

4.1.1 Compensatory Principle

There are three major principles to allocate functions of human and machines and design human-machine systems. The Compensatory Principle (Hollnagel and Bye, 2000; Hollnagel and Woods, 2006) is one of them, in which functions of humans and machines are separated and allocated to what they are good at. The origin of this principle is known as Fitt's MABA-MABA (what "Men Are Better At" and what "Machines Are Better At") list (Fitts, 1951) shown in Table 4.1, and Hollnagel and Bye (2000) characterized it as the Compensatory Principle.

Table 4.1: MABA-MABA list (Fitts, 1951; Inagaki, 2003).

Humans appear to surpass present-day machines with respect to the following:
1. Ability to detect small amounts of visual or acoustic energy.
2. Ability to perceive patterns of light or sound.
3. Ability to improvise and use flexible procedures.
4. Ability to store very large amounts of information for long periods and to recall relevant facts at the appropriate time.
5. Ability to reason inductively.
6. Ability to exercise judgment.
Present-day (in 1950s) machines appear to surpass humans with respect to the following:
1. Ability to respond quickly to control signals and to apply great forces smoothly and precisely.
2. Ability to perform repetitive, routine tasks.
3. Ability to store information briefly and then to erase it completely.
4. Ability to reason deductively, including computational ability.
5. Ability to handle highly complex operations, i.e., to do many different things at once.

This principle aims at avoiding excessive demands to human performance and ensuring efficient human-machine interactions. However, their roles often tend to be so fixed that any flexibilities of their operations could be lost in real fields of practice. It is, therefore, necessary for this principle to take into account the validity of such static function allocation in ever changing environment, for which no practical means had not been established yet.

The following case study addresses this issue by investigating an actual air crash accident with one of the FRAM simulation models. Specifically, the FRAM simulation envisions how the validity of the operation, from the perspective of the feasibility of operational procedures, had been changed by variabilities of working environment generated in the accident sequence. This case study consequently provides an example to demonstrate the importance of such validity examinations in this principle, or stress test of artifacts, including not only physical objects but also intangible products such as manuals or checklists, against their surrounding environment.

4.1.2 Aviation: One of the Most Advancing Fields of Automation

The aviation field has a long history of the automation, through which they have a variety of experiences and lessons about their operations as already described in the chapter 2. In this context, the design and feasibility of Standard Operational Procedures (SOPs) are often discussed for their safety. Kirlik (1993), for example, discussed about the complexity of the procedures for using automation in the flight-deck and concluded based on his experiment that most of the pilots do not use automation aids in the crowded air space. This is because pilots heavily stressed to carry out the flight-deck procedures for changing automation settings when the change of heading and altitude is requested by Air Traffic Controller (ATC). Degani and Wiener (1994) also investigated the design of flight-deck procedures, involving three major U.S. airlines. They carried out several activities such as interviewing people working for the airlines, attending procedure design meetings, and jump seat observations. Through the research, they found several cases in which the deviation from SOPs occurred for the sake of efficient or comfortable flight. The research is concluded that there is nothing such as an optimal set of procedures nor "royal road" to procedure development, and the room for individualism remains for the execution of SOPs even in highly proceduralized system. They suggest that the feasibility of SOPs is not stable; the feasibility of SOPs could change depending on situations, and it is essential for us to understand the characteristics of the unstable feasibility.

4.2 Case Study: Safety Analysis of Aviation Flight-Deck Procedure

This case study is an investigation of actual air crash accident that occurred near Cali Airport, Colombia in 1995. This was the first fatal accident of the high-tech B757 aircraft in its 13 years of exemplary service at that time. Its direct trigger is assumed to be an incorrect entry of approach course to the Flight Management Computer (FMC) by the pilots and subsequent loss of their situation awareness; a complicated design of a HMI and SOPs for the autopilot are considered to be the contributing factors as well. It is even said that the accident would not have happened if it were not for the automation.

4.2.1 Overview of the Accident

American Airlines flight 965 was about to land at an airport close to Cali, Colombia. The flight 965 was originally supposed to fly over the airport to south and then turn towards north for final approach during the landing sequence. In this context, an Air Traffic Controller (ATC) proposed a runway change to the flight 965 since the ATC assumed everything (e.g., weather or traffic conditions) was fine to make straight in. The pilots of the flight 965 accepted this proposal because the schedule was already two hours behind due to a departure delay at Miami, and it was already dark outside; this change of the flight plan seemed to be rational to everyone involved in this case. However, the pilots suddenly became busy in responding to the new flight plan after this. Specifically, the shortened approach course brought about a number of tasks that are to be performed in a very limited available time (i.e., under an extremely high time pressure). They were consequently confused with identifying the new approach course which is called Rozo One Arrival in the end; the Rozo One Arrival requires aircraft to overfly TULUA before proceeding to ROZO as shown in Fig. 4.1.

The pilots finally identified this course after a while. However, it was too late because the flight 965 had already flown over the TULUA — the initial point of the Rozo One Arrival. They got confused again with how to fly over the TULUA for initiating the Rozo One Arrival and then came up with an alternative plan, i.e., flying direct to the ROZO; this is evidently an irregular operation, but no one, including even the ATC, noticed and warned about this.

One of the pilots eventually entered information into the Flight Management Computer (FMC) in order to fly towards the ROZO. Specifically, he just entered

"R" instead of "ROZO" into the FMC and executed the command without cross-checking. However, the FMC interpreted it as an input cord of "ROMEO" of Bogota, Columbia beginning with "R" in the same way. This is because the HMI of the FMC was designed to display a waypoint which is frequently used rather than the nearest at the top of the list if there are multiple waypoints starting with same characters. The pilots did not notice the mistake and entrusted the computer to make a course change. In the end, the flight 965 swerved off the original course and crashed into mountainous terrain existing in the east of the airport as shown in Fig. 4.1.

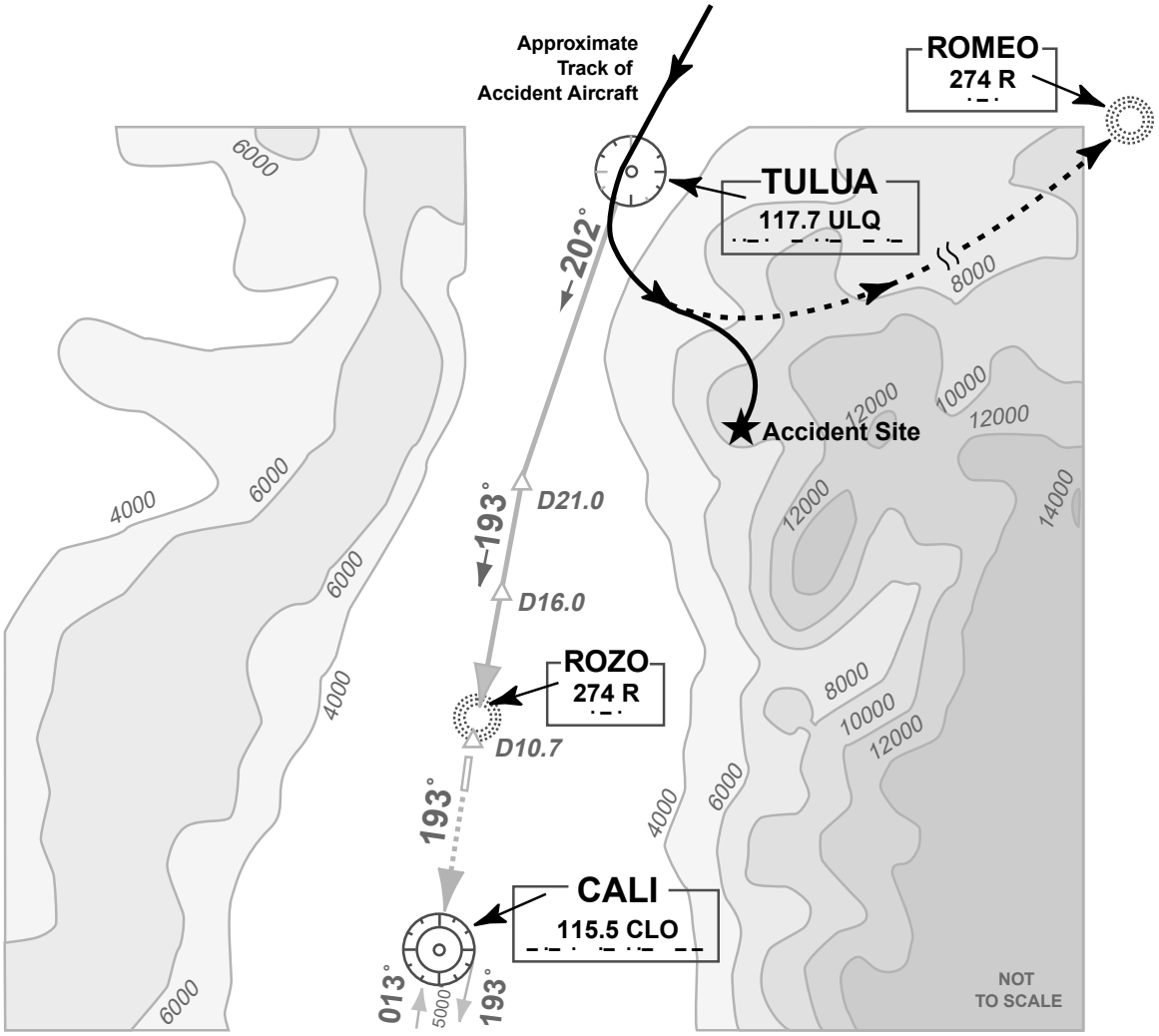


Fig. 4.1: Rozo One Arrival and estimated track of the flight 965 (Simmon, 1998).

4.2.2 Initial Setting for Analysis

One of the most critical points of this accident is that the pilots of the flight 965 tried the irregular operation and entered the wrong course into FMC. In addition, although

runway change is basically regarded as a normal event in daily operation, it seems to have caused the fatal error of the pilots in this case. The focus of this analysis is therefore put on what took place until they entered the wrong course into the FMC. Based on this, the parameters required for analysis are defined as following.

Functions and Their Potential Couplings

The first step of the analysis is to identify functions of FRAM, and we identified five functions as shown in Table 4.2 here. Note that the identification number of each functions in Table 4.2 have nothing to do with the order of execution; the order or dependencies among functions are determined for the first time when the potential couplings are instantiated, and an instance is obtained. Besides, the potential couplings among functions are identified as shown in Fig. 4.2; they are instantiated before the calculation process and generate various instances.

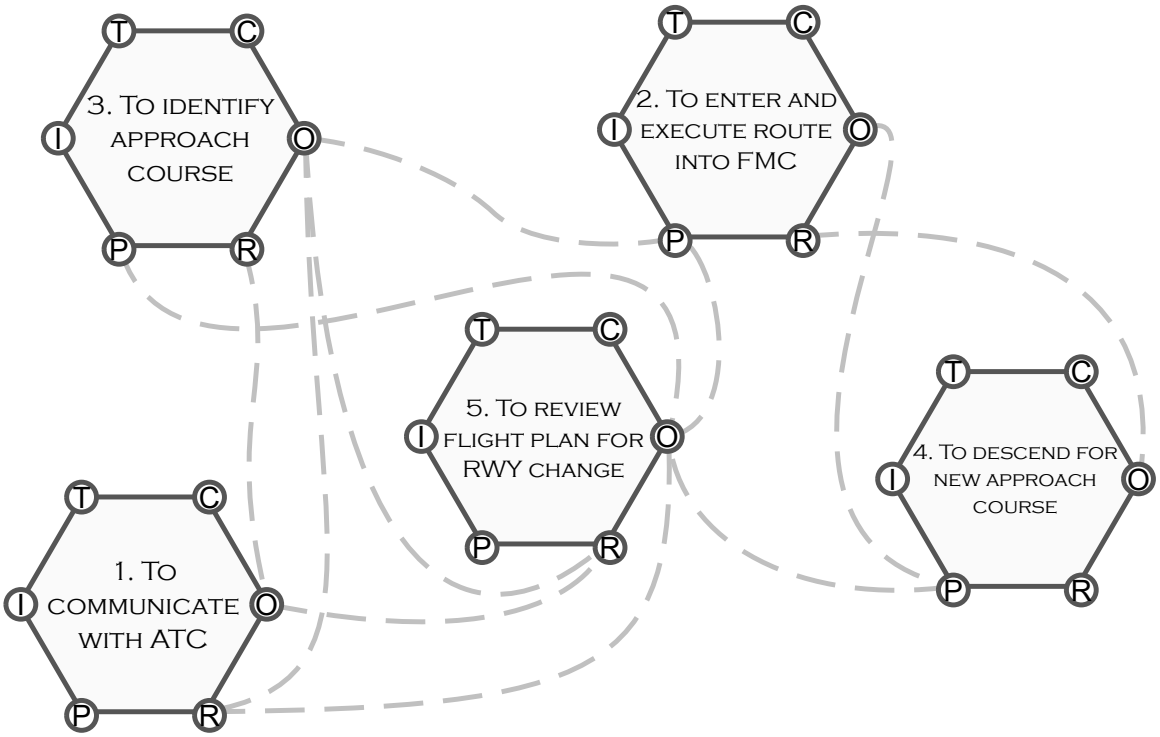


Fig. 4.2: All potential couplings between functions.

Table 4.2: Functions which were required to be performed after the flight 965 received the proposal of runway change.

1. To communicate with ATC		2. To enter and execute route into FMC	
Input	N/A (Not applicable)	Input	N/A
Output	Communication is established correctly	Output	Aircraft starts to fly correct course
Precondition	Frequency is set correctly	Precondition	Correct course is identified
Resource	Information exchanged among crews and ATC	Resource	Information from chart, instrument and ATC
Control	Crews, ATC, radio equipment, radar	Control	Crews, FMC, ND: Navigation Display
Time	Several tens of seconds	Time	Several tens of seconds
3. To identifying approach course		4. To descend for new approach course	
Input	N/A	Input	N/A
Output	Approach course is identified	Output	Flight continues correctly
Precondition	Crew recognize the current flight status	Precondition	New flight plan is validated
Resource	Information from chart, instrument and ATC	Resource	Altitude, distance (Available time)
Control	Crews, ATC	Control	Pilot Flying (PF)
Time	Several tens of second	Time	Few minutes
5. To review flight plan for RWY change			
Input	Accepting runway change		
Output	Flight plan is validated		
Precondition	Crew recognize current flight status		
Resource	Altitude, distance (Available time)		
Control	Crews, ATC		
Time	Several tens of seconds or few minutes		

Weight of CPCs

The next step is to define relative weight of CPCs in each function as shown in Table 4.3. In this analysis, they were defined with following procedure:

1. Identify CPCs which have the most significant effects on the performance of a subject function.
2. Let the relative weight of those CPCs, i.e., W_i in Eq. 3.2 be 100.
3. Evaluate the relative weight of other CPCs based on the above evaluation.

It should be noted that these values are normalized automatically with Eq. 3.2 in the implemented process of the simulation.

Table 4.3: CPC weight in each function.

CPC	Functions				
	1	2	3	4	5
Availability of resource	50	20	100	100	100
Adequacy of training and experience	10	10	20	50	20
Quality of communication	100	5	30	0	100
Adequacy of MMI	0	100	10	10	10
Access to procedures	10	100	10	10	10
Working condition	40	30	30	100	60
Number of goals	50	60	80	100	60
Available time	50	60	80	100	60
Circadian rhythm	10	5	5	5	10
Crew collaboration	20	100	50	20	100
Organization factor	5	0	0	0	5

Simulation Scenario

The last step of the initial setting is to represent variabilities occurred in the accident sequence with manual operations of the CPC scores. To this end, we identified the following five major variabilities, in addition to the initial state, based on a prior investigation (Simmon, 1998) and examined their effect on all instances generated from potential couplings shown in Fig. 4.2.

> *Initial state*: Everything was supposed to be going well. Therefore, all CPC scores of all functions are assumed to be 100 at the beginning of the accident sequence.

Variability 1: There were continuous discrepancy of communication after the flight 965 entered the control area of the Cali airport. This reduced the score of "Quality of communication" in To communicate with ATC to 20.

Variability 2: The prior investigation (Simmon, 1998) points out that the runway change was accepted without sufficient review of the flight plan. This reduced the score of "Crew collaboration quality" in To review flight plan for RWY change to 0.

Variability 3: After the acceptance of the proposal, one of the pilots became busy in manual operation for descending. This reduced the score of "Available time" and "Number of simultaneous goals" in To descend for new approach course to 0.

Variability 4: The pilots of the flight 965 had to identify a new approach course under high time pressure. This reduced the score of "Available time" and "Number of simultaneous goals" in To identifying approach course to 0.

Variability 5: The prior investigation (Simmon, 1998) points out that the letter "R" which is the initial letter of next waypoint was input without cross-checking. Also, the letter "R" was programmed as a completely different place from their destination at that time. Then, the score of "Adequacy of HMI" and "Crew collaboration quality" in To enter and execute route into FMC is set to 0.

4.2.3 Simulation Results

In this simulation, changes of $log(PAF)$ in each function were obtained with respect to all generated instances. Among them, we specifically put focus on two instances since all of the instances were basically generated systematically, and it is irrational to take into account all of them; some of them are not reasonable to execute as a procedure. The one is considered to represent the procedure executed during the accident sequence. The other is picked up by analyst because it showed a quite different result from the former one even though the structure of those instances are

so alike each other, implying that the safety of socio-technical systems has something to do with the design of operational procedures.

The former instance is shown in Fig. 4.3, and the transition of $log(PAF)$ in its functions are shown in Table 4.4 and Fig. 4.4; the numbers on abscissa in Fig. 4.4 correspond to that of *Variability 1 – 5*. According to Table 4.4 and Fig. 4.4, the $log(PAF)$ of all the functions increased after *Variability 1*, which means the situation became dangerous. In addition, while the $log(PAF)$ in almost all the functions remained constant under the influence of *Variability 2 – 5*, the $log(PAF)$ of To enter and execute route into FMC increased to -0.70 due to *Variability 5*. This value indicates that the control mode of this function is "Scrambled" which is the most dangerous state, according to Table 3.6.

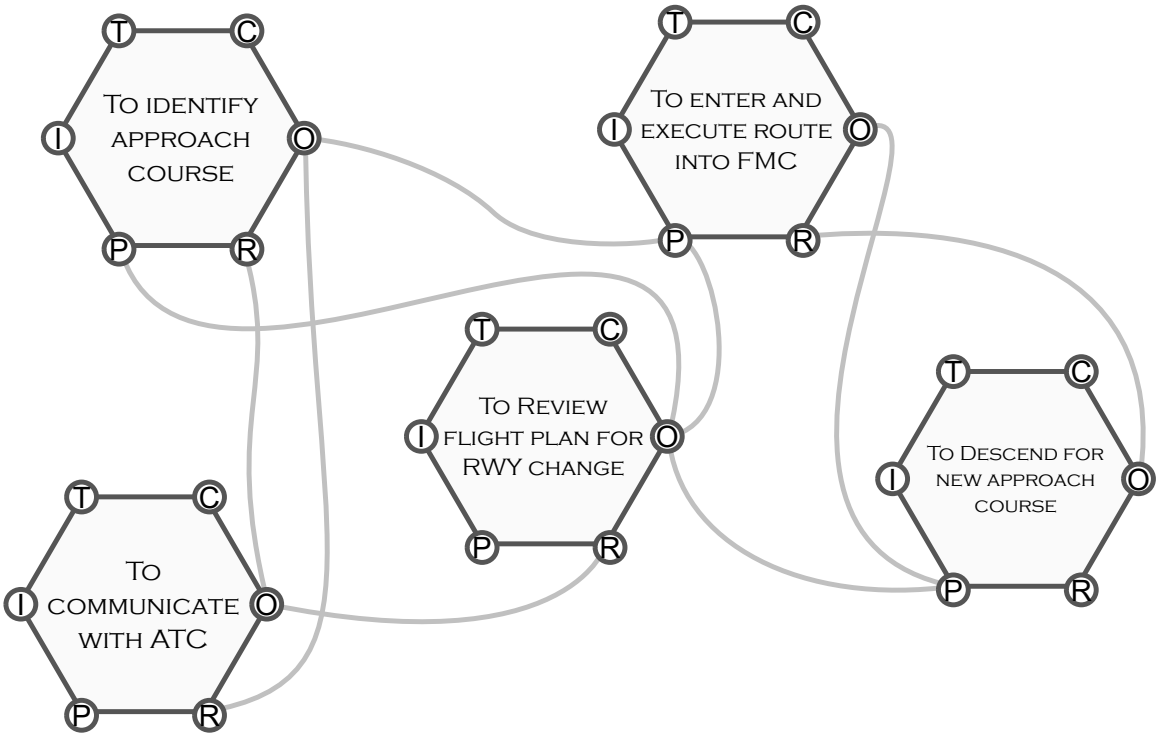


Fig. 4.3: Instance which became unsafe due to variabilities.

Table 4.4: Transition of $log(PAF)$ in each function in Fig. 4.3.

Function ID Number	*Variability*					
	Initial State	1	2	3	4	5
1	-4.80	-1.61	-1.61	-1.61	-1.61	-1.61
2	-4.80	-2.68	-2.69	-2.69	-2.69	-0.70
3	-4.80	-1.60	-1.61	-1.61	-1.60	-1.60
4	-4.80	-1.65	-1.68	-1.61	-1.61	-1.62
5	-4.80	-1.60	-1.60	-1.60	-1.60	-1.60

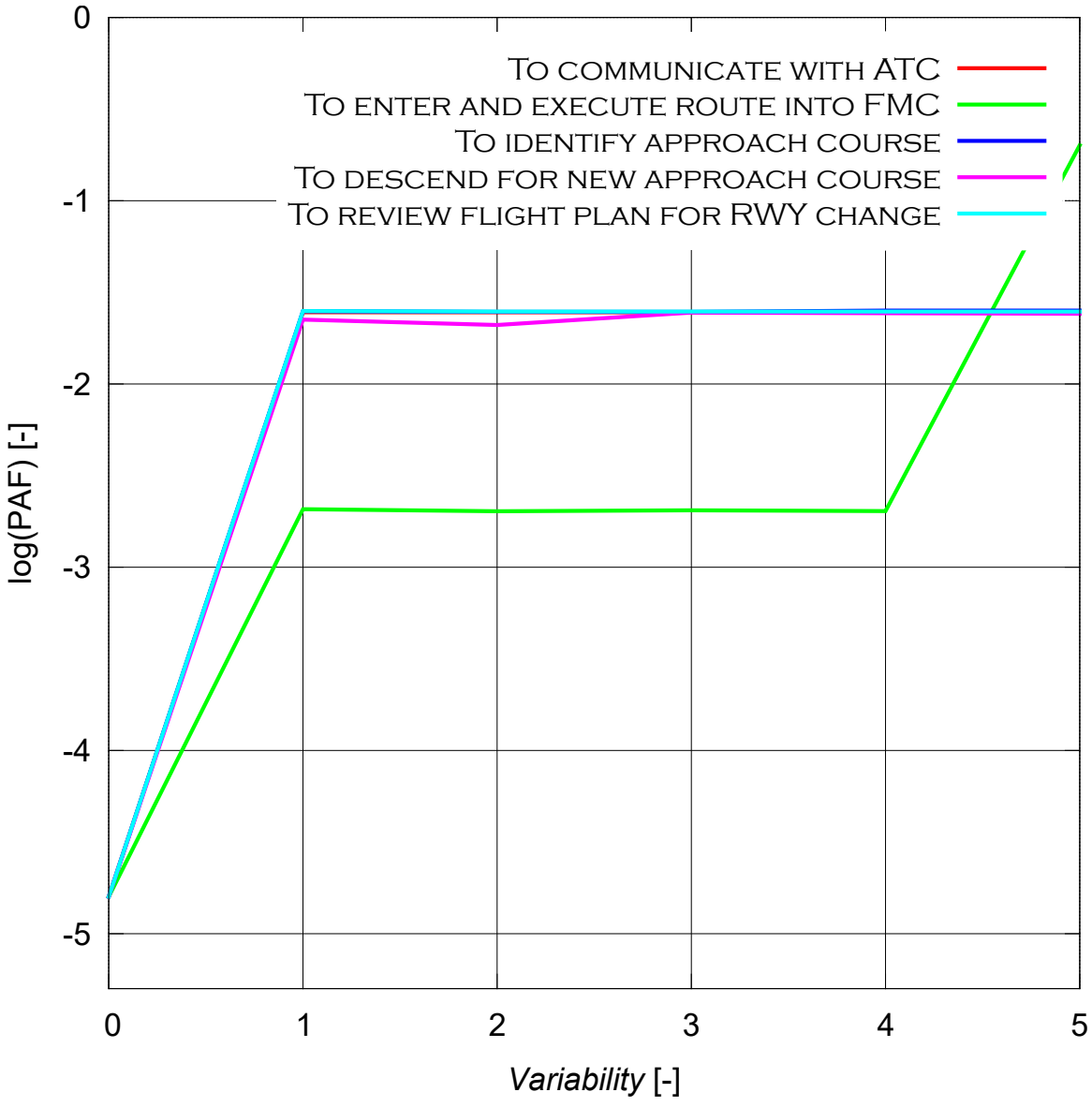


Fig. 4.4: Graphical representation of Table 4.4.

The latter instance is shown in Fig. 4.5, and the change of $log(PAF)$ in its functions are shown in Table 4.5 and Fig. 4.6 as well as the previous case. Compared to the previous case, their behaviors were quite different. Specifically, the $log(PAF)$ in To enter and execute route into FMC was -1.62 after *Variability 5*, which was lower than that the case of Fig. 4.4. In addition, the $log(PAF)$ in To descend for new approach course started decreasing after *Variability 3*, and it reached -2.91 at the end of this simulation, whose control mode corresponds to "Tactical," according to Table 3.6. Therefore, this instance can be regarded as safer than that in Fig. 4.3.

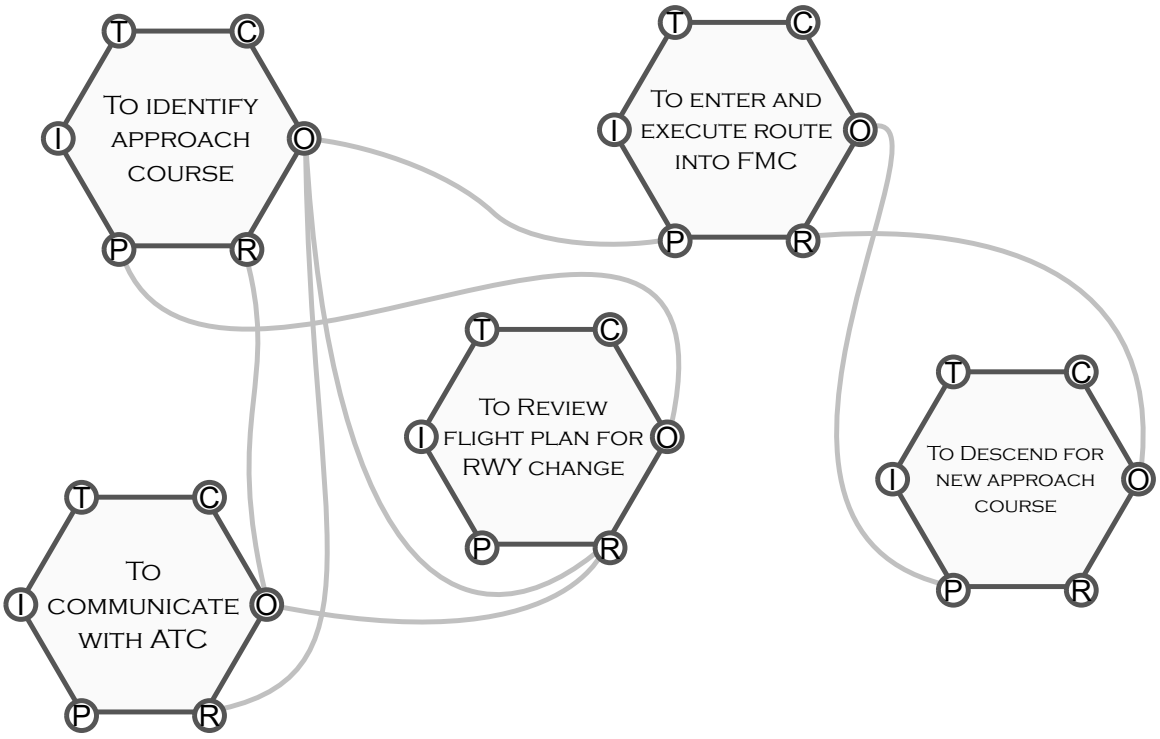


Fig. 4.5: Instance which remained safe against variabilities.

Table 4.5: Transition of $log(PAF)$ of each function in Fig. 4.5.

Function ID Number	*Variability*					
	Initial State	1	2	3	4	5
1	-4.80	-1.61	-1.61	-1.61	-1.61	-1.61
2	-4.80	-2.85	-2.86	-2.77	-2.79	-1.62
3	-4.80	-1.61	-1.61	-1.61	-1.60	-1.60
4	-4.80	-3.99	-3.99	-2.76	-2.77	-2.91
5	-4.80	-1.62	-1.62	-1.62	-1.62	-1.62

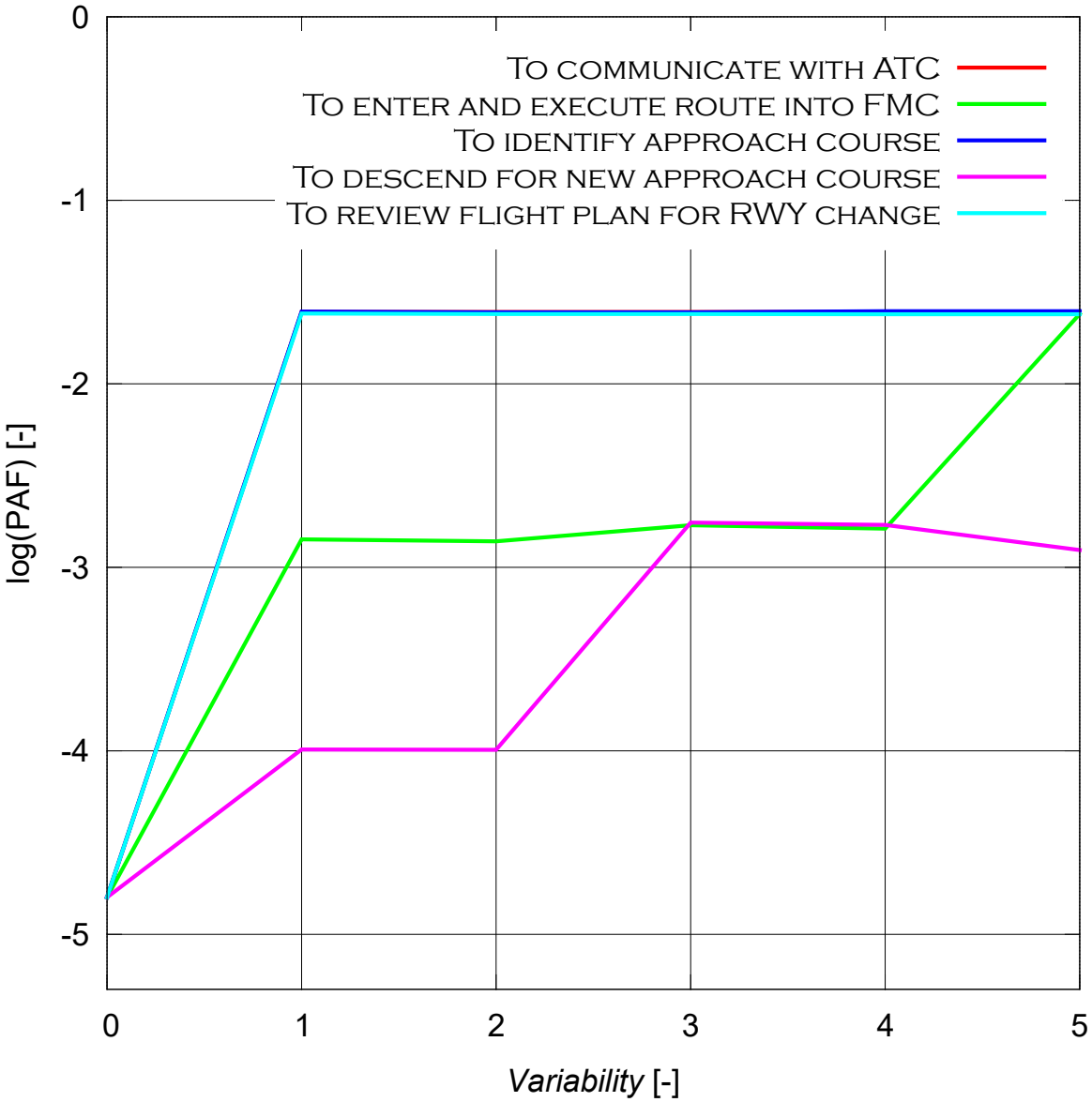


Fig. 4.6: Graphical representation of Table 4.5.

4.2.4 Discussion

According to the result, the behavior of each instance against the same variabilities were quite different from each other, depending on their structures. Compared to the instance in Fig. 4.3, the output of To review flight plan for RWY change in Fig. 4.5 has only one connection with the other function, which makes a major difference between actual procedures represented by Fig. 4.3 and Fig. 4.5. This means that the order to execute functions in Fig. 4.5 is more explicit than that in Fig. 4.3, and the explicit order of executions plays an important role to make the

procedure feasible in the context of this accident. Eventually, the simulation result suggests that the safety of socio-technical systems could vary depending on the design of operational procedures, and catastrophes could be avoided with the proper design of procedures in adversities.

It should be noted that *Variability 1 – 5*, that are manual operations to change CPC scores, subsequently caused the resonance among CPCs, variabilities in the functions, i.e. change of PAF, and the resonance among functions automatically. Moreover, the dependency among CPCs or functions, which cannot be controlled arbitrary, played a significant role in this simulation, and the above result cannot be obtained without such dependencies. This kind of result is peculiar to this FRAM model and cannot be obtained with other safety analysis methods.

Another approach to analyze socio-technical systems has been provided in the field of social science, for instance. The theory is called *activity theory* originating from the psychology studied by Vygotsky (1980) and developed by Engeström (1987). In this theory, activities are understood to have a mutually linked structure, in which the three factors (communities, rules, and division of labor) are added in addition to human being, object, and mediating tools, to trace the alternation and development of activities as "a succession of contradictions" generated in this structure and among multiple activities. Safety analyses based on this kind of idea have focused on unsafe acts such as human errors and/or breach of rules by front-line workers; these unsafe acts directly affect the system safety and are characterized by that the negative impacts of such acts become conspicuous relatively quickly (active failure).

On the contrary, there are latent factors, of which impacts do not become apparent quickly and lay hidden without bringing any harm, but afterward can destroy the system defense by interacting with the local environment (latent condition). People who work within a complex system commit unsafe acts, such as errors or breach of rules, by some reasons that cannot be explained with the psychology studying an individual. Such reasons continue to be hidden within organizations and never become apparent unless facing abnormal conditions resulted from the unsafe acts or others. Such latent conditions do not lay behind statically but change with time by interacting each other below the surface. The processes of such changes may lead to changes in front-line works, as well as to accidents in the worst scenario. For instance, empirical evidences of the criticality accident occurred at the nuclear fuel conversion facility of JCO are reported (IAEA, 1999). This accident was caused by the intentional transformation of work procedures by onsite workers. This transformation was driven by the various contradictions, i.e., misfits between components of the

activity system, might arise out of and propagate throughout their activities, then induce some sorts of changes in the procedures for the better or for the worse. Such procedure transformation is dealt with as adaptation of procedures, and this work can contribute to visualizing the process of its changing with latent contradictions, thus contribute to the in-depth analysis of organizational accidents.

4.3 Future Prospects to Utilize Proposed FRAM Simulation

4.3.1 Design of Operational Procedures

In future work, we consider the design of operational procedures as an effective way to practice the resilience engineering with the proposed model. What is important for resilient systems is to maintain validity of systems in a specific context where they are operated. In this respect, the simulation result in the previous section suggested that the safety of the systems could vary depending on the design of procedures in a specific context. Therefore, one goal to make the systems resilient is to find out proper actual works for safe operations in specific contexts, and we call it, in this context, "design of operational procedures".

The process is to modify predetermined procedures with elaborate validation as shown in Fig. 4.7. To validate the design of procedure, we use the proposed FRAM simulator as a tool of stress test of procedures in a specific context. The simulator can evaluate the feasibility of procedures in a specific context which can be represented by the score of CPCs, and this stress test can contribute to finding out proper actual works. Here, it should be noted that this process is intended to evaluate the validity of evolving socio-technical systems shown in Fig. 2.9.

4.3.2 Safety Analysis based on Safety-II

From the perspective of the resilience engineering and Safety-II, not only why things go wrong but also why things go right must be taken into account. Specifically in this context, we should additionally learn from good practices in which original operational procedures were violated by the execution of some non-regulated actions allowing any "further catastrophe" to be avoided; the case known as the "Miracle on the Hudson" in 2009 was one of these cases (NTSB, 2010), for example. However, lessons learned so far have mainly focused on risks and been deduced from an analysis of failures that

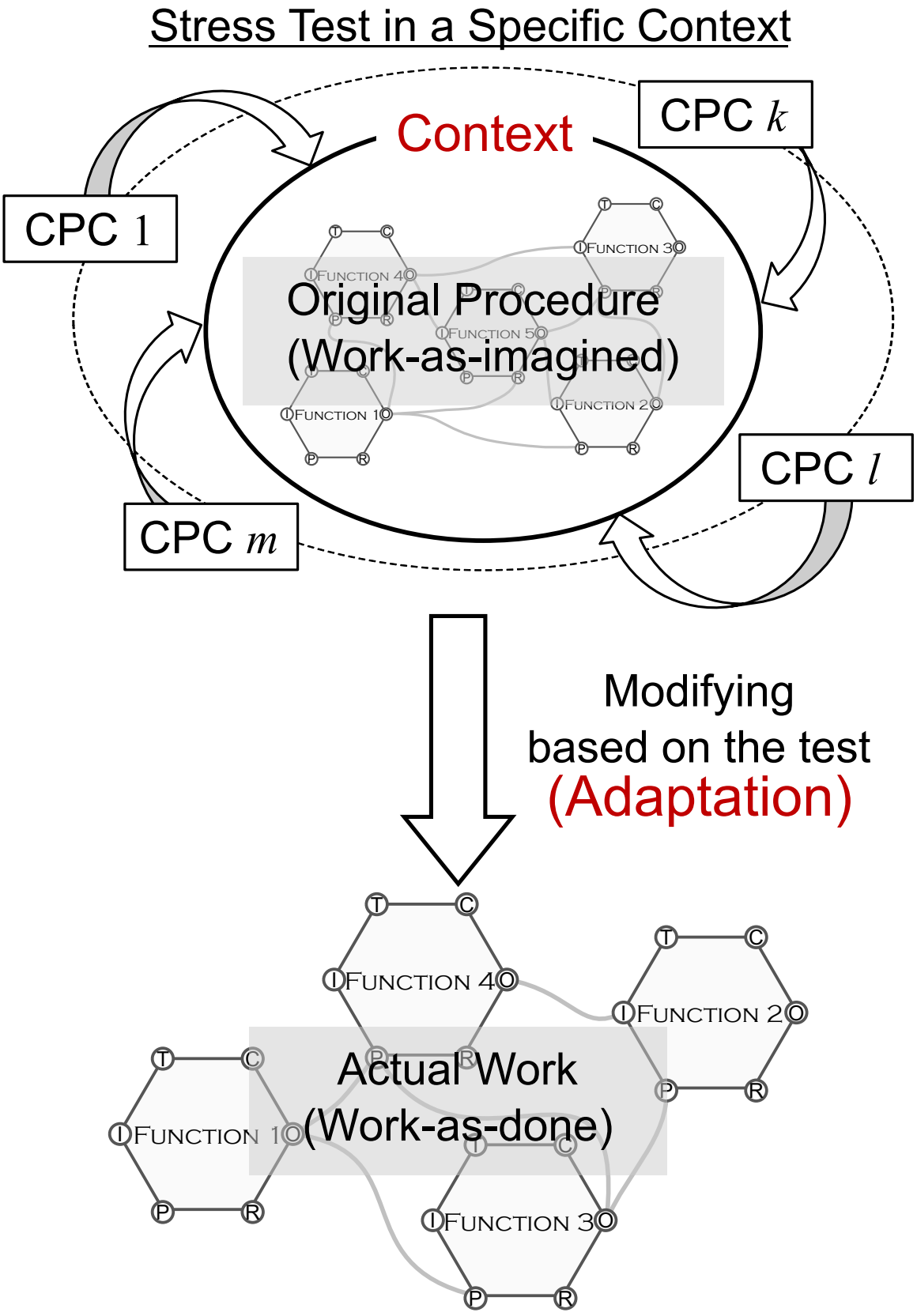


Fig. 4.7: Design of operation procedures in a specific context.

led to accidents. In spite of the existence of potential lessons to be learned, there has been almost no analysis, assessment nor lessons gleaned with respect to such cases.

The proposed FRAM simulator could contribute to this purpose; the FRAM can evaluate validity of actual works and give us clues to figure out why they are working well in a particular context. This contributes to extending a particular success to a more generalized lesson that are valid even in a changing environment and provides new lessons to improve the capability to handle "unforeseen contingencies." The knowledge obtained from this visualization could especially contribute to adaptation or modification of procedures in the design of operational procedures. Finally, we accumulate the result of those simulations which contain the authorized knowledge about actual works and generalize it for the proper design of operational procedures. This consequently contribute to the cross-industrial enhancement of the safety.

4.4 Brief Summary

In this case study, the simulation has envisioned a vulnerability of the Compensatory Principle and provided insights about how to cope with it. Although this kind of issues have already been discussed since before, it has inevitably been too qualitative or conceptual, as the limitation of the functional safety indicates. In this respect, the FRAM model has provided insights/criteria to support and facilitate the discussions as the certain simulation results here. This is a great progress from the traditional approaches, and the FRAM model has enabled us to extend the domain of analysis.

Chapter 5

Implicit Role of Human Beings Imposed by Leftover Principle

5.1 Introduction

5.1.1 Leftover Principle

The Leftover Principle (Hollnagel and Bye, 2000; Hollnagel and Woods, 2006) is another principle to design human-machine systems. This principle is to automate everything found to be feasible, and its purpose is to ensure process efficiency of operations. However, this also means that human beings are pushed to a domain where too complex tasks/activities to be automated are left and need to cope with all of them. Operations of systems based on this principle shall, therefore, depend on implicit knowledge and skills of human operators, and it is generally difficult to justify their validity.

This problem is now regarded as critical especially in the manufacturing industry in Japan. This is because those who have this kind of knowledge and skills are retiring due to their age with increasing its speed; their knowledge and skills must be inherited to the next generations, otherwise the industry, which is one of the greatest advantages of this nation, could easily decline. Therefore, the nature of such implicit knowledge/skills must be elucidated for sustainable development of the industry, and moreover, the nation.

The following case study is to examine one of such empirical knowledge currently inherited in the steel production industry. Specifically, the FRAM simulation envisions how such knowledge works in complex dynamics of the system's operation and provides insights for its justification.

5.1.2 Complexity of Steel Production: Difficulties to Anticipate Its Operations

Steel production typically uses a kind of socio-technical system, and the operations involved are often too complex to anticipate. The steel industry is generally said to contain far more uncertainties than other industries, and it is difficult to predict operations such as delivery period or production load (Shioya et al., 2015). This is mainly due to the number of product variants with which the steel industry must contend; their production processes highly depend on the specifications of those products as well as on their quality. In addition, the processes are affected by quantities of in-process inventory, system troubles, and maintenance, all of which can be related to human operators, machines, organization, and the working environment. These issues combine to make the process of steel production too complex to be anticipated.

To build a simulation model for this process, we first investigated empirical knowledge. We found that proper adjustment of the rate of direct delivery can improve material flows of a supply chain in actual steel production, according to an engineer working at a steel production company. Here, the rate of direct delivery is the ratio of what is sent directly to a downstream production process to what is sent to a storage space in the production line. However, no one can systematically explain why the adjustment is valid; how the adjustment works is so complex that no one can follow the process. The primary purpose of this case study is to investigate the mechanism and determine the features of such emergent outcomes.

f

5.2 Initial Setting of Simulation

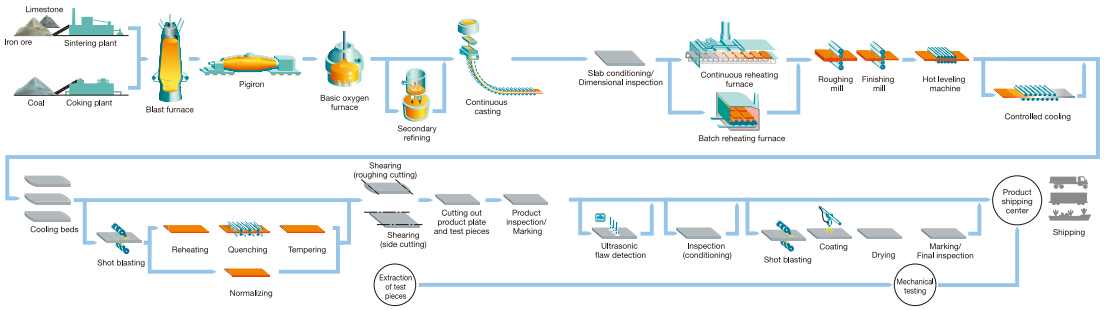
In this simulation, the second — cellular automaton based FRAM model is adopted to investigate dynamic and continuous behavior of the target system. The simulation starts with the initial setting shown in Fig. 3.14, which needs to be done manually. It includes settings for both functions and CPCs. We also need to come up with a simulation scenario. The details are as follows.

5.2.1 Functions and Their Dependency

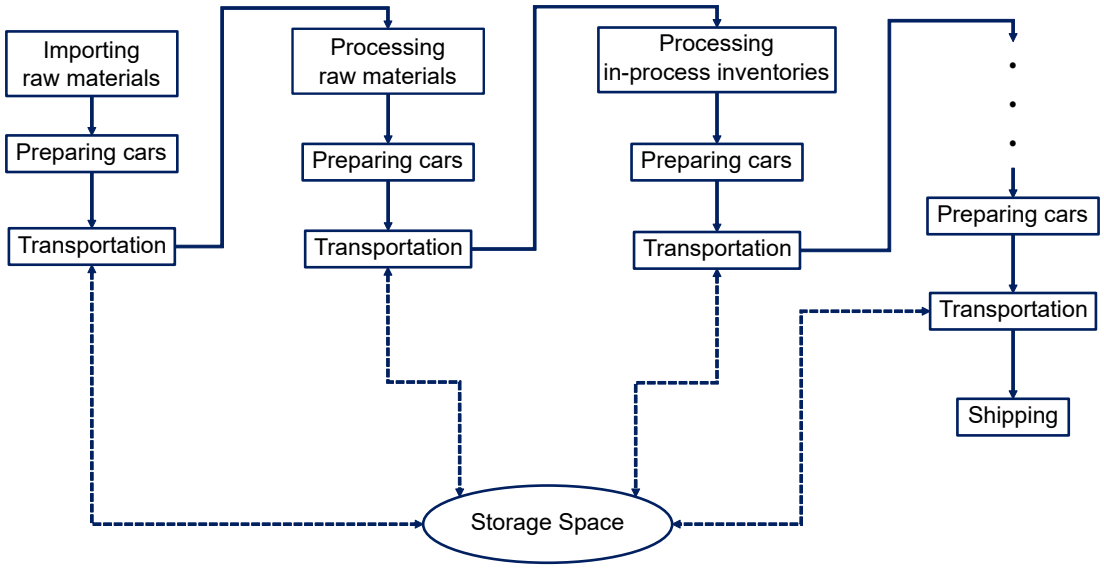
Figure 5.1(a) shows the target system of this simulation, which is a typical production process of steel plates. The starting point of the process is importing raw materials. These materials go through production processes such as steel making, continuous

casting, and rolling. While they are transported among the processes, some materials, in-process inventories, and products are temporarily sent to storage spaces.

The processes in Fig. 5.1(a) are converted into a more abstracted configuration to determine the FRAM functions, and Fig. 5.1(b) shows the physical dependency of each abstracted process. This provides five FRAM functions first of all: To import raw materials, To prepare cars, To transport, To process raw materials/in-process inventories, and To ship. We also consider two new functions: To send to storage space and To prepare storage space, which are related to the storage space and the dotted lines in Fig. 5.1(b). Therefore, a total of seven functions were defined in this simulation, with their details listed in Table 5.1. In addition, their potential coupling was defined as shown in Fig. 5.2, which is used as an instance in this simulation. It should be noted that some of the function names in Table 5.1 and Fig. 5.2 are slightly different from those in Fig. 5.1(b) for convenience.



(a) Actual production process of steel plates (STEEL, 2019)



(b) Abstracted representation of steel production process

Fig. 5.1: Typical production process of steel plates.

Table 5.1: Functions to produce steel plates.

1. To import raw materials		2. To transport among processes	
Input	Not Applicable (N/A)	Input	Raw materials are ready
			In-process inventories are processed
Output	Raw materials are ready	Output	Transportation is ongoing/done
Precondition	N/A	Precondition	Cars are ready
Resource	N/A	Resource	Cars
Control	N/A	Control	N/A
Time	N/A	Time	N/A
3. To Process		4. To prepare storage space	
Input	Transportation is done	Input	In-process inventories are processed
Output	In-process inventories are processed	Output	Storage space is prepared
Precondition	N/A	Precondition	N/A
Resource	In-process inventories	Resource	Stored in-process inventories/products
Control	N/A	Control	Flow rate of in-process inventories/products
Time	N/A	Time	N/A
5. To prepare cars		6. To send to storage space	
Input	Raw materials are ready	Input	Transportation is ongoing
	In-process inventories are processed		
	Storage space is prepared		
Output	Cars are ready	Output	In-process inventories/products are sent to storage space
Precondition	N/A	Precondition	Storage space is prepared
Resource	N/A	Resource	Storage space
Control	Flow rate of in-process inventories/products	Control	Flow rate of in-process inventories/products
Time	N/A	Time	N/A
7. To Ship			
Input	Transportation is done		
Output	N/A		
Precondition	All processes have successfully completed		
Resource	N/A		
Control	N/A		
Time	N/A		

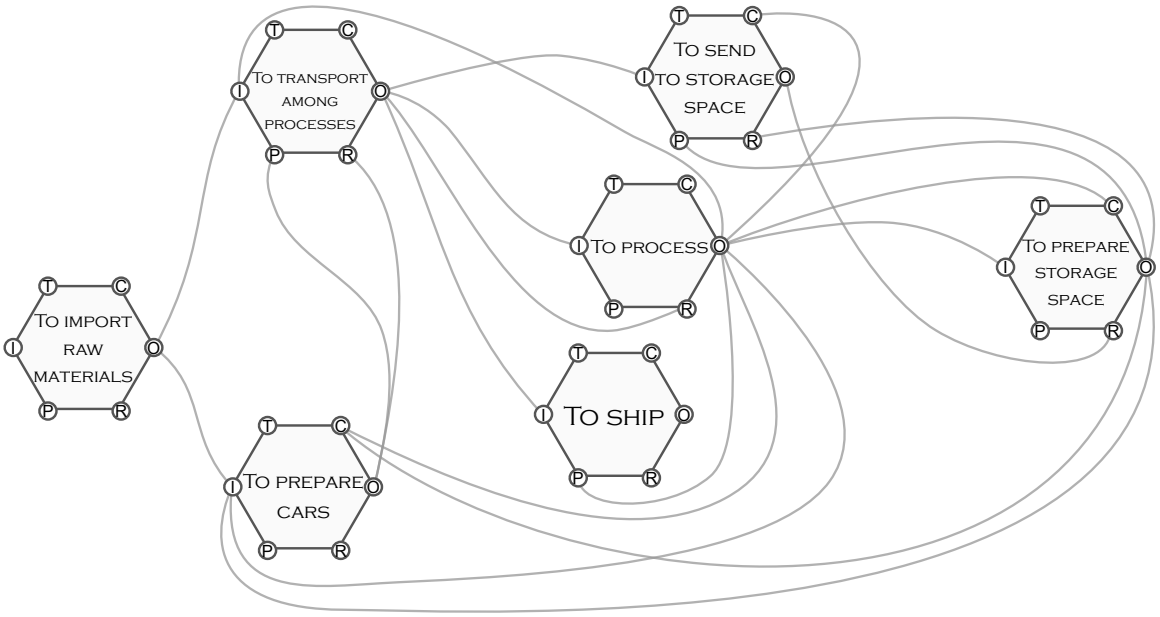


Fig. 5.2: Dependency among functions: instance of steel production.

5.2.2 CPC Belonging to Each Function

All functions shown in Table 5.1 and Fig. 5.2 are assumed to have a set of CPCs. In this simulation, a new set of CPCs is introduced for more specialized simulation of steel production. Specifically, the four original CPCs — Training and experience, Man-Machine interaction, Circadian rhythm, and Organization factor — were replaced with "Quality of materials," "Adequacy of lot size," "Timeliness," and "Adequacy of direct delivery rate." The four new CPCs are closely related to the operation of steel production and have a significant effect on it, according to an engineer working for a steel production company, while the four original CPCs seems trivial or much less significant. Besides, too many CPCs can make the simulation too complicated, leading to the improper setting of initial parameters or to results that do not make sense. In total, eleven new CPCs were introduced in this simulation, which are shown in Table 5.2; their weights are shown in Table 5.3 in which the identification numbers of functions and CPCs correspond to that of the functions and CPCs in Table 5.1, and 5.2, respectively.

Table 5.2: Dependency among the new set of CPCs.

	Original CPCs	Replaced with
1	Available resources	
2	Adequacy of training and experience	Quality of materials
3	Quality of communication	
4	Adequacy of Human-Machine Interface	Adequacy of lot size
5	Availability of procedures	
6	Working conditions	
7	Number of simultaneous goals	
8	Available time	
9	Circadian rhythm	Timeliness
10	Crew collaboration quality	
11	Adequacy of organization	Adequacy of direct delivery rate

Table 5.3: CPC weight in each function.

	Functions						
CPCs	1	2	3	4	5	6	7
1	0.0588	0.217	0.153	0.245	0.224	0.189	0.224
2	0.176	0.0191	0.153	0.0225	0.0190	0.0166	0.0815
3	0.0588	0.0191	0.0508	0.0225	0.0815	0.0406	0.0249
4	0.176	0.0470	0.0508	0.0980	0.0815	0.0966	0.0815
5	0.0196	0.0191	0.0169	0.0225	0.0190	0.0166	0.0190
6	0.1765	0.0470	0.0508	0.0980	0.0249	0.189	0.0815
7	0.0196	0.0470	0.0508	0.0273	0.0815	0.0406	0.0815
8	0.0588	0.217	0.153	0.0980	0.0815	0.0406	0.0815
9	0.176	0.217	0.153	0.245	0.0815	0.142	0.224
10	0.0196	0.0470	0.1525	0.0225	0.224	0.0406	0.0815
11	0.0588	0.105	0.0169	0.0980	0.0815	0.189	0.0190

The CPC weights were defined using a paired comparison process that is a part of the analytic hierarchy process (AHP) (Saaty, 1990). This is because the CPC weight is based on qualitative relationships between a FRAM function and its related CPCs, and it would involve too much subjectivity if it were defined intuitively. AHP was originally a methodology to support decision making (e.g., buying a car) on the basis of multiple criteria (e.g., price, fuel efficiency, size). In the process, one paired comparison of those criteria is carried out. The relative importance among each pair is evaluated with integer grades, and they are regarded as the ratio of one criterion

weight to the other. Those ratios yield the following matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} \frac{w_1}{w_1} & \frac{w_1}{w_2} & \cdots & \frac{w_1}{w_m} \\ \frac{w_2}{w_1} & \frac{w_2}{w_2} & \cdots & \frac{w_2}{w_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{w_m}{w_1} & \frac{w_m}{w_2} & \cdots & \frac{w_m}{w_m} \end{bmatrix} \tag{5.1}$$

where w_i $(1 \leq i \leq m)$ is the weight of the i-th criterion and m is the number of criteria. The following vector is a set of weight to extract from the above matrix:

$$\mathbf{w} = \begin{bmatrix} w_1, w_2, \cdots, w_m \end{bmatrix}^T. \tag{5.2}$$

The multiplication of the matrix in Eq. 5.1 and the vector in Eq. 5.2 yields

$$\begin{aligned} \mathbf{Aw} &= \begin{bmatrix} \frac{w_1}{w_1} & \frac{w_1}{w_2} & \cdots & \frac{w_1}{w_m} \\ \frac{w_2}{w_1} & \frac{w_2}{w_2} & \cdots & \frac{w_2}{w_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{w_m}{w_1} & \frac{w_m}{w_2} & \cdots & \frac{w_m}{w_m} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix} \\ &= m \begin{bmatrix} w_1, w_2, \cdots, w_m \end{bmatrix}^T \\ \iff & \quad \mathbf{Aw} = m\mathbf{w}. \end{aligned} \tag{5.3}$$

Equation 5.3 suggests that the set of CPC weights can be obtained by solving the eigenvalue problem, and the set of weights corresponds to an eigenvector whose eigenvalue is the closest to m — the number of criteria. In this simulation, the criteria correspond to the weight of CPCs; the above pair-wise comparison process was used to obtain the sets of CPC weight in Table 5.3 for each function.

In addition to the weight of CPCs, dependency among CPCs needs to be defined. The dependency among original CPCs has already been provided in Table 3.3. However, the new set of CPCs shown in Table 5.3 was introduced in this simulation, which is why the dependency among them needs to be redefined. This is shown in Table 5.4, where CPCs in the rows are affected by the CPC in the columns; the identification numbers in the first column correspond to those in the second row. Here, 0 means there is no dependency among two corresponding CPCs, and 1 means the CPC in the column has an effect on the CPC in the row.

Table 5.4: Dependency among the new set of CPCs.

	CPCs which have effect on										
CPCs which are affected by	1	2	3	4	5	6	7	8	9	10	11
1. Available resource	0	1	1	1	1	1	1	1	1	1	1
2. Quality of materials	1	0	0	0	0	0	1	1	1	1	0
3. Quality of communication	1	0	0	0	0	1	1	1	1	1	0
4. Adequacy of lot size	1	1	1	0	0	0	0	1	1	1	1
5. Availability of procedures	1	0	1	0	0	0	0	0	0	1	0
6. Working condition	1	0	1	0	0	0	0	1	1	0	0
7. Number of simultaneous goals	1	1	0	1	1	1	0	0	0	1	1
8. Available time	1	1	0	1	1	1	1	0	1	1	1
9. Timeliness	1	1	1	1	1	1	1	1	0	1	0
10. Crew collaboration	1	0	1	1	1	1	0	1	0	0	0
11. Adequacy of direct delivery rate	1	1	1	1	0	0	0	0	1	1	0

5.2.3 Simulation Scenario

A simulation scenario needs to be set and converted into a manual change of parameters, such as CPC scores or dependency among functions. The scenario was set as shown below in this simulation.

Scenario: There was an excess arrival of raw materials, and the flow of materials, in-process inventories, and products grew beyond the capacity of the steel production processes. As a means for overcoming the adversity, the rate of direct delivery was adjusted at a specific timing.

This is converted into manual changes of FRAM entities, referred to as *Variability* and *Countermeasure*:

Variability: The score of CPC "Available resource" of the function To import raw materials is set to 0, which corresponds to the worst state of the CPC, at simulation time $T = 0$.

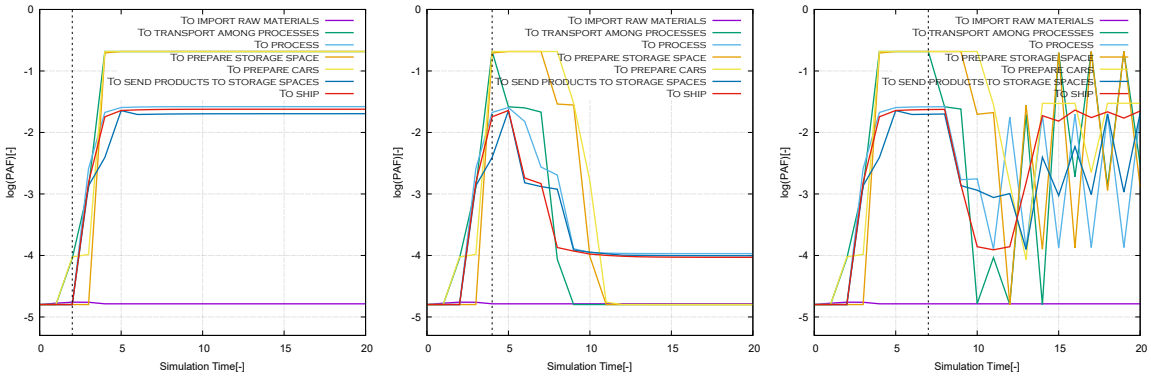
Countermeasure: The simulation process is paused once at a specific timing, and the score of CPC "Adequacy of direct delivery rate" of the function To Transport among processes is set to 100, which corresponds to the best state of the CPC.

In this simulation, every timing to take a *Countermeasure* is collectively examined to determine the difference of each case. One hundred transition patterns of $log(PAF)$ for each function is obtained as a result.

5.3 Simulation Results

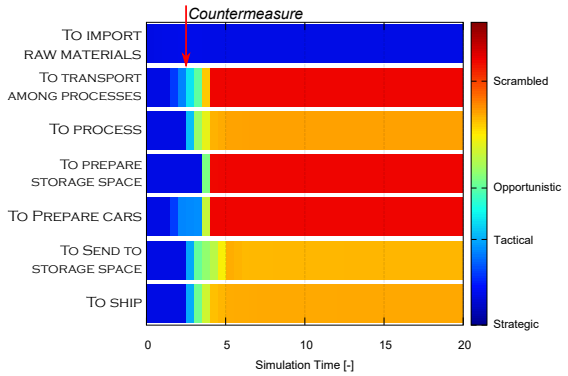
Figure 5.3 shows three characteristic transition patterns of the $log(PAF)$ of each function. The horizontal and vertical axes represent the simulation time and the $log(PAF)$ of each function, respectively. The $log(PAF)$ is equivalent to the degree of danger/instability of each function. The dotted line in the figure represents the timing when *Countermeasure* was adopted. For example, in the case of Fig. 5.3(b), $log(PAF)$ and control mode of To transport among processes was initially -4.80 and Strategic, respectively at $T = 0$. Then, they had been degraded to -0.68 and Scrambled by *Variability* until $T = 4$. *Countermeasure* was taken at this time, and the $log(PAF)$ and control mode recovered to -4.80 and Strategic, respectively again. What is interesting with these results is that they were explicitly classified into only the three patterns shown in Fig. 5.3, even though 100 transition patterns of $log(PAF)$ were obtained.

It should be noted here again that the simulation results shown in Fig.5.3 are dynamical transitions of the safeties of the functions, whose plotted values are defuzzified $log(PAF)$s according to Eq. 3.7. Since these values were defined initially as fuzzy variables, it would be plausible to reinterpret the dynamical safety status qualitatively (i.e., in fuzzy linguistic values) at the end. Figure 5.4 represents the transitions of the control modes of each function qualitatively with color gradations. The relationships between the values of $log(PAF)$ shown in Fig. 5.3 and the values of the control mode are determined according to the definitions of the fuzzy values shown in Table 3.6. Hereafter, the simulation results are mainly analyzed on the basis of what Fig. 5.4 represents, and Fig. 5.3 is also referred to if necessary.

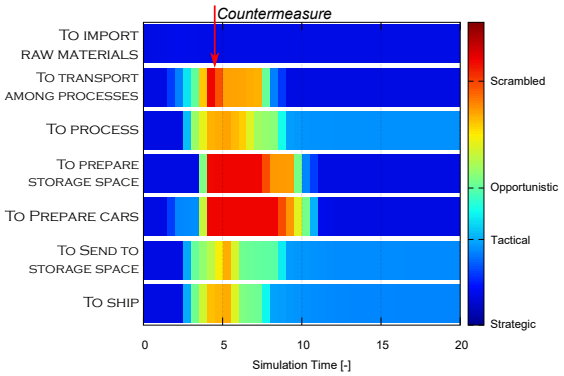


(a) Pattern 1: *Countermeasure* was taken earlier than $T = 4$ (b) Pattern 2: *Countermeasure* was taken at $T = 4$ or $T = 5$ (c) Pattern 3: *Countermeasure* was taken later than $T = 5$

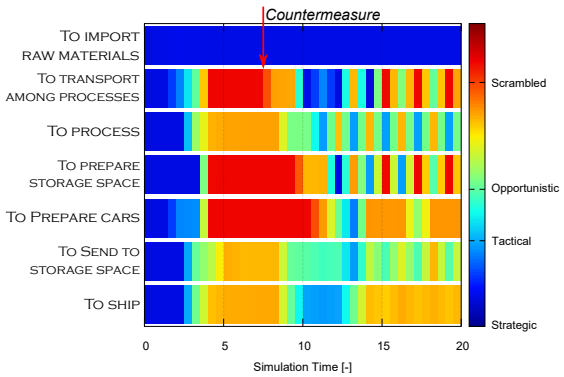
Fig. 5.3: Simulation results: Transition patterns of $log(PAF)$ in each function.



(a) Qualitative representation of Pattern 1 shown in Fig. 5.3(a)



(b) Qualitative representation of Pattern 2 shown in Fig. 5.3(b)



(c) Qualitative representation of Pattern 3 shown in Fig. 5.3(c)

Fig. 5.4: Qualitative representation of simulation results: transition of control mode in each function.

5.3.1 Interpretation of Simulation Results

The pattern in Fig. 5.4(a) shows the transitions of the control mode in each function when *Countermeasure* was adopted earlier than simulation time $T = 4$. In this case, the control modes of each function, except To import raw materials, started their transitions towards undesirable states after *Variability* at simulation time $T = 0$; they were expected to be calmed down by *Countermeasure*. However, they did not go back to the original state even after *Countermeasure* was adopted, and as such, retained an unsafe status. Moreover, the control modes of To transport among processes, To prepare storage space, and To prepare cars revealed at the end the most dangerous state of the four control modes: Scrambled. This designates a situation in which *Countermeasure* failed in preventing the occurrence of the growing danger/instability of the functions, resulting in the most dangerous/unstable state of the three patterns.

The pattern in Fig. 5.4(b) shows a case when *Countermeasure* was taken at simulation time $T = 4$ or $T = 5$. In this case, all of them went back to the safe state, Strategic or Tactical, autonomously after *Countermeasure* was adopted, in contrast to the previous case. It is remarkable here that the impact of *Countermeasure* recovered the control mode of the rest of the functions indirectly, even though *Countermeasure* itself was originally intended to improve only the control mode of To transport among functions. This pattern implies a situation where *Countermeasure* can work very effectively, and the system can go back to the safe state.

The pattern in Fig. 5.4(c) shows a case when *Countermeasure* is adopted later than simulation time $T = 5$. This case is characteristic in the sense that the control mode of all functions except To import raw materials and To prepare cars started revealing periodic transitions, even oscillations, after *Countermeasure*, and some of them were interrelated with each other. For example, the recovery of To process led the control mode of To transport among processes to a dangerous/unstable state, and vice versa; a similar trend can be seen in the relationship between To prepare storage space and To send storage space, as well. Moreover, the control mode of To prepare cars was degraded again and remained at Opportunistic, closer to Scrambled, after *Countermeasure* was adopted, meaning that it was difficult to prepare cars and that the resources for transportation were strictly limited. These transition patterns reveal a situation in which the recovery of one function resulted in increasing its outputs (e.g., raw materials, in-process inventories, and products) while adversely resulted in too much provision for the other functions due to the lack of cars. In other words, there were trade-off relationships among the pairs of functions because the resource for transportation was strictly limited, thus causing the oscillation patterns.

5.3.2 Factors of Difference Among Three Patterns

The factor behind the difference between Pattern 1 (Fig. 5.4(a)) and Pattern 2 (Fig. 5.4(b)) can be found in the control mode of To transport among processes before and after *Countermeasure* was adopted. In the case of Pattern 2, *Countermeasure* succeeded in recovering the control mode of To transport among processes directly, and its effect indirectly recovered the control mode of the rest of the functions as well. On the other hand, *Countermeasure* failed to recover the control mode of To transport among processes in Pattern 1, as if the "power" of the control mode to become worse was so strong that *Countermeasure* could not calm it down, resulting in the most undesirable state of the three patterns. The "power" in this context

stemmed from the dependency among CPCs formulated as Eq. 3.9, since *Countermeasure*, or manual change of the CPC score, did not bring about any changes of the control mode in Fig. 5.4(a), implying that the effect of *Countermeasure* vanished. These two patterns suggest that the functions cannot recover without the success of *Countermeasure*, and that it becomes effective after simulation time $T = 4$, when the control mode of To transport among processes reaches the most unstable state, Scrambled, and stops getting worse. In other words, the difference between Pattern 1 and Pattern 2 depends on whether *Countermeasure* can recover the control mode of To transport among processes or not, and it is useless to adopt *Countermeasure* earlier than the above turning point.

Similarly, the factor behind the difference between Pattern 2 (Fig. 5.4(b)) and Pattern 3 (Fig. 5.4(c)) can be found in the recovery process of the control mode of To process. The recovery speed of the control mode after *Countermeasure* in the pattern of Fig. 5.4(c) was slower than that of Fig. 5.4(b), and it prevented the other functions from recovering. This is mainly because the $log(PAF)$ of To process shown in Fig. 5.3(c) was gradually increasing until just before *Countermeasure* was adopted; the increase was so small that it could not be observed as a change of the control mode shown in Fig. 5.4(c). In the end, minor changes in $log(PAF)$ or the control mode of To process caused the major difference between Pattern 2 and Pattern 3.

5.3.3 Summary of Simulation Results

The results ultimately suggest that the effect of *Countermeasure* becomes the most significant when the control mode of its target function, To transport among processes, gets as tense or unstable as the effect of *Variability* can cause. In other words, it is most effective to wait until the state of the target function becomes the most tensed or unstable state, according to Pattern 1 and Pattern 2. On the other hand, the safety of To process is also degraded little by little while waiting, and it could cause the unstable outcomes shown in Fig. 5.4(c) as well. Therefore, we conclude that there is a critical timing when *Countermeasure* can become the most efficient, as shown in Fig. 5.4(b), and *Countermeasure* at any other timing can cause chaotic outcomes, as shown in Figs. 5.4(a) and 5.4(c).

5.4 Discussion: Using Complexity for the Safety of Socio-Technical Systems with the Proposed Model and Its Future Prospects

There are two remarkable findings in the obtained result. One is that, even though *Countermeasure* itself was only intended to directly improve the safety of To transport among processes, its impact indirectly improved the safety of the rest of the functions as well. The other is that *Countermeasure* became effective just after its target function, To transport among processes, had gotten as tense or unstable as the effect of *Variability* could make it. There has been some discussion regarding how to predict the future evolution of complex systems and manage them accordingly, and these characteristics are somewhat related to topics that are controversial in those discussions.

5.4.1 First Finding: Efficient Control of Complex Systems

While it might be ideal to have every artifact system, including socio-technical systems, under control, it is almost impossible to do so since socio-technical systems involve many factors in their operation, requiring infinite precision. We can find some clues to overcome this problem in the following. Israeli and Goldenfeld (2004) have suggested that computationally irreducible physical processes can be computationally reducible at other coarse-grained resolution levels. In other words, it is possible to predict the behavior of complex systems without accounting for all of their small-scale details. Also, Smith and Johnson (2004) contended that it is not necessary to know all about target systems; only minimal knowledge and intervention on certain parts of the systems is enough to manage their future evolution. These studies convincingly argue that it is not required to understand everything about a target complex system nor to intervene on a large part of the systems in order to control it. Instead, it is sufficient to only intervene on a small, specific part of the systems — one whose effect interacts with the other parts according to its potential dynamics and leads to desired outcomes. Moreover, complex systems can be harnessed by taking advantage of their characteristics of complexity and emergence.

The transition patterns of the control modes of the functions shown in Fig. 5.4(b) precisely imply the above assertions. There is no need to intervene in the CPC scores of multiple functions to overcome the adversity caused by *Variability*. Instead, it would be enough to adjust the CPC score of just one particular function: "Adequacy of direct delivery rate" in To transport among processes. In other words, the

impact of a small intervention would propagate all around the FRAM network shown in Fig. 5.2, interacting with the rest of network and bringing about the most efficient and desirable outcome compared with the other two cases shown in Fig. 5.4. While experienced engineers in the steel production industry do know this fact unconsciously, at the same time, they also know that it is not always effective, as shown in Fig. 5.4. Thus, to help them overcome this confusion, tools that enable such complex, possible behaviors to be envisioned are required. This is precisely the goal of the work, and our proposed model does contribute to deliberating on the means for the indirect control (i.e., harnessing) and to clarifying why things go right. It brings to life the concept of Safety-II, thus enhancing the resilience of socio-technical systems.

5.4.2 Second Finding: Drastic Change of Complex Systems

Complex systems can change their behavior drastically under certain circumstances (Mitchell, 2009; Johnson, 2009), which is generally called phase transition or bifurcation. The particular circumstances in which a phase transition occurs are called critical points, where small changes in specific parameters can cause qualitative changes in the behavior of a macroscopic system (e.g., between water and vapor at boiling point). One of the most famous examples of this phenomenon can be seen in a flock of birds: the entire flock can change its velocity and flying direction drastically, making it difficult for predators to attack. However, this does not mean that all the birds in the flock always fly in the same direction at the same speed; the parameters of each bird usually fluctuate and correlate with each other, causing information flow through the flock network. Bialek et al. (2014) investigated this mechanism by building a model based on the maximum entropy method, which succeeded in fully representing data observed in the real world. Their research implies that a system with high entropy could be close to a critical point, and that it can exist between order and chaos — known as the edge of chaos.

Figure 5.3 reveals a similar phenomenon to the above phase transition. When the transition from Pattern 1 to Pattern 2 in Fig. 5.3 is taking place just after the control mode of To transport among processes, the target function of *Countermeasure* reaches its peak of $log(PAF)$, where the instability of this system is assumed to be high. Also, such an unstable state contributed to the transition from Pattern 2 to Pattern 3: the transition is caused by the minimal increase in the $log(PAF)$ of To process around simulation time $T = 5$. That is, just small changes in specific parameters could cause drastic changes of system behaviors, thus demonstrating critical points or the edge of chaos. Similar behavior can be observed

in the transition patterns of the control mode in Fig. 5.4. Ultimately, the drastic changes might also lead to undesirable outcomes, and well-considered management is essential to deal with such a situation.

Although it might sound conflicting with our intuitions: adversities or unwanted outcomes must be avoided as soon as possible. There indeed exist these kinds of gaps between theory and practice in discussions about safety, and they are frontiers to explore for the future safety of socio-technical systems. What is essential to manage the safety of socio-technical systems is to consider, understand, and harness such complex dynamics of systems. The FRAM model can provide us with insights about new safety methodologies, i.e., resilience engineering based on Safety-II.

5.4.3 Limitations and Future Improvements of This Model

There are three major points to be improved on this FRAM model. They are closely related to both reliability and efficiency of the simulation and must be addressed in future works. On the other hand, it is also sure that those improvements bring about great impacts on this FRAM model, making it much more attractive.

The first point is that the current FRAM model requires many parameters for its initial setting of the simulation, which might consume too much energy, time, and involve subjectivities. It is especially case with the setting of CPCs because their weight needs to be defined for all functions, and the process is now depending on analysts. Also, a set of CPCs can be customized as shown in this simulation, which is also pointed by Konstandinidou et al. (2006) and Zhou et al. (2017a). That is why we need to seek for some ways to collectively and efficiently define factors of working environment as CPCs, depending on subjects of the simulation. Besides, the weight of those CPCs needs to be evaluated, based on objective and automatic solutions. One possibility to solve this problem is to apply machine learning techniques to the above process in which the model is fed by real data and analyze it to create a set of CPCs and their weight. If this process is automated, it can improve both the efficiency and reliability of this FRAM model.

The second point is that the simulation results currently just provide abstract or conceptual insights about the safety of socio-technical systems. In other words, those results just provide qualitative state of the target systems, and it is difficult to know quantitative information about the safety. To address the issue, we are now assuming that another less abstracted simulations such as physical simulations are required, if necessary, to investigate what will actually happen in the real world. Moreover, the simulation of FRAM running at a higher abstraction level and some

physical simulations running at a lower abstraction level need to be connected in some ways. One solution for the purpose is extracting dynamical transitions of CPC scores changing behind $log(PAF)$ in each FRAM function. This is because they can involve more detailed information about a working environment and provide semi-quantitative criteria about the result of simulation. In other words, they can work as if they were media connecting between the higher and lower abstraction levels and bring about input and output values for both of the simulations.

The last point is that it requires some experience to interpret the results since they are still too conceptual. In this simulation, the original results of Fig. 5.3 were converted into qualitative color maps to provide a better understanding. However, we still need to seek other ways to represent those results with more simplified symbols so that the systems' status can be grasped at a glance. This will be important, primarily when the simulation is used in real operations in letting the operators know the status of systems immediately. For this purpose, we are currently expecting that "Force Dynamics" (Talmy, 1988) can be useful for visualizing such dynamics The Force Dynamics is a semantic category to describe how entities interact concerning force, including the exertion of force, resistance to such a force, and the overcoming of such a resistance. According to this, the following three patterns are distinguished as different dynamical patterns:

Pattern 1: The growing effect of variability propagating through the system is currently so strong that the effect of countermeasure is hidden, and the system cannot recover from the chaotic status.

Pattern 2: The growing effect of variability propagating through the system became weaker and weaker, and this makes countermeasure effective; thus the system can overcome the variability turning out to be stable.

Pattern 3: The effect of countermeasure propagating through the system can contribute to recovering only a part of the system, while it, in turn, makes the rest of the system still in danger.

In this way, the Force Dynamics is a useful scheme for the qualitative envisioning of the overall system status and is expecting to transfer the qualitative representations shown in Fig. 5.4 onto more abstract summarization contributing to the more intuitive understanding of the results.

5.5 Brief Summary

In this simulation, we investigated the effect of empirical knowledge of engineers in the steel production industry by adjusting the rate of direct delivery. The results suggested that some characteristics of complex systems — namely, harnessing, phase transformation, and the relationship between critical point and entropy — played a significant role in the dynamics. This means that the human beings in the Leftover Principle are implicitly demanded to cope with such complexities, and the fact has been envisioned as a result of simulation. This is just an individual case at this time, and the result needs to be generalized as future work.

Chapter 6

Justification of Complementarity/ Congruence Principle for Future Safety of Artifacts

6.1 Introduction

6.1.1 Complementarity/Congruence Principle

The Complementarity/Congruence Principle (Hollnagel and Woods, 2006) is the last principle to design human-machine systems. The focus of this principle is put on how to enhance coagency based on active interactions between human and machines, support their long-term comprehensions, and maintain their control under a variety of conditions. The importance to realize this kind of human-machine relationships is rapidly increasing for future safety of artifacts.

The speed of technological change is currently accelerating, and highly sophisticated technologies, such as Artificial Intelligence (AI), Internet of Things (IoT), and advanced automations driven by them, are coming close to our daily life. These technologies are expected to automate our daily lives (e.g., houseworks or driving cars) more and more, where a variety Levels of Automation (LoA) is mixed. To ensure the safety, it is essential to consider how to realize effective human-machine collaborations based on this principle, not just clearly distinguishing roles of human and machines and leaving tasks/activities to the automation as much as possible. This is because the history shows that the automations without effective human-machine collaborations often result in confusions of human beings and catastrophic consequences; the situation could be even worse since the consumers are generally "novices" of the automations, contrary to well-trained "professionals" such as aviation pilots. However, this point of view seems to be missing in current Research and Development (R&D)

of such technologies in many cases; those technologies are sometimes designed as if to say their purpose is to reduce the workload of human beings, and human-machine collaborations are therefore not necessarily important.

The following case study is to suggest this is not true. This case study examines the feasibility of the SAE conditional driving automation in time-critical situations. Specifically, the FRAM simulation investigates how to realize smooth authority transfer from the autonomous driving system to human drivers in an emergency; the result provides an evidence to support the importance of human-machine collaborations.

6.1.2 SAE Conditional Driving Automation

The autonomous driving technology is currently developed for commercial use. We can classify the autonomous driving technologies developed so far into one of the five Levels of Driving Automation (LoDA) issued by SAE (2016). According to the definition, the system is basically responsible for the Dynamic Driving Tasks (DDTs) above the LoDA 3. The Research and Development (R&D) of the technology is currently aiming at the LoDA 3 or conditional driving automation, where human drivers still have to intervene in the DDTs in an emergency.

Table 6.1: Each Levels of Driving Automation (LoDA) and their definitions (SAE, 2016; Inagaki and Sheridan, 2019).

LoDA	Definition
1. Driver assistance	Automation performs either longitudinal or lateral vehicle motion control (on a sustained basis), but not complete Object and Event Detection and Response (OEDR)
2. Partial driving automation	Automation performs both longitudinal and lateral vehicle motion control (on a sustained basis), but not complete OEDR
3. Conditional driving automation	Automation performs the complete Dynamic Driving Task (DDT), but not DDT fallback, within a limited Operational Design Domain (ODD)
4. High driving automation	Automation performs the complete DDT and DDT fallback within a limited ODD
5. Full driving automation	Automation performs the complete DDT and DDT fallback without ODD limitation

The authority transfer from the autonomous driving system to the drivers in an emergency has primarily been one of the most challenging issues in this context. Many researchers in both academic and industrial fields have been working on the subject. Their primary interests are on how many seconds are required to request the drivers to take over the authority and disengage the automation. However, it is still on the way to find out the fundamental solution; their proposals often sound so ad-hoc that they can be valid to just limited situations.

In this regard, Inagaki and Sheridan (2019) and Eriksson and Stanton (2017) point out the context-dependency of conditions to transfer the driving authority. Inagaki (2003) points out that the conventional discussions about function allocation between human operators just focus on "who does what," and the perspective of "who does what and WHEN" is missing there; the function allocation should be adaptive to the situations where operational functions should be shared or traded between the human operators and systems, depending on a specific context. Eriksson and Stanton (2017) reviewed 25 kinds of researches about the responding time to the Take Over Request (TOR) to point out that the previous researches put too much focus on the average time to issue the TOR and take over the authority. They also concluded that the mean or median values do not tell the entire story of the authority transfer through their additional experiment, and the strategy of the authority transfer should be adaptive to situations.

Despite the nature of this problem, the current context of the R&D for the autonomous driving technology is eager to find out the static or even "fixed" time for the drivers to smoothly take over the authority in an emergency. This inconsistency is now leading us to the never-ending exploitation and away from the fundamental solutions. The problem is that it is quite difficult to properly envision actual situations and practical requirements in the R&D process.

This is exactly the nature of the envisioned world problem, suggesting that the current R&D of the autonomous driving technology is conducted within this framework. It is, therefore, necessary to leave the physical or real-oriented approaches once and carry out investigations at a higher — functional abstraction level to overcome this problem. In this case study, the third — complex adaptive systems based FRAM model is applied for the purpose of it.

6.2 Overview and Objective

This simulation is to investigate the safety of the authority transfer problem in time-critical situations. The crucial difference from the traditional approaches is that this simulation does not consider the real, physical conditions such as a distance to an obstacle (e.g., construction site) ahead of the vehicle. Besides, it neither deals with the exact time needed before the issue of TOR informing when the automation will be disengaged. Instead, this case study is to model the functional safety of the operation of the Autonomous Driving System (ADS) and investigate how its safety is resilient against the variabilities to address the envisioned world problem.

Especially the objective of this simulation is to investigate the feasibility of conditional driving automation at LoDA 3, where the human driver should still be the final authority of the driving and must take evasive actions in an emergency. In such situations, the driver is required to process good Situation Awareness (SA) and make consequent Decision Making (DM) when the TOR is issued, and the ADS inevitably has to provide the appropriate cognitive support with the human driver. According to Endsley (1995a), the SA consists of the three levels shown in Table 2.2. The formal definition of SA is broken down into three segments: perception of the elements in the environment, comprehension of the situation, and projection of future status. In order to make a smooth authority transfer from automation to a human driver, the systems should also guarantee supports that are appropriate for each of the three segments of a driver's SA. That is, a human driver, whom the TOR is issued and prepares for the authority transfer from automation, has to attain successful SA conducted by the support of the system. In an emergency, it would be ideal for supporting all segments of SA thoroughly as earlier as possible, but it would not always be possible. Its success would depend upon which cognitive status a driver is and upon what support is provided by the system. Complex behaviors may emerge wherein, and the functional simulation is needed to find out the best human-automation collaboration.

6.3 Initial Setting of Simulation

Some parameters should be set to start the simulation as well as other case studies. The first item of the initial setting is to model the target system with the functions. The second item is to set the weight of CPCs for each function. We also need to come up with a simulation scenario and describe it with parameters of the FRAM model. The details are shown below.

6.3.1 Functions and Their Potential Couplings

The operation of autonomous driving is modeled as shown in Fig. 6.1, and the detail description of the functions are presented in Table 6.2 in this simulation. The model was built based on fundamental tasks of the drivers, including SA, DM, and their manual operations. Further details about each function are described below. Note that the potential couplings are currently illustrated with dotted lines since they just represent the possible dependency among functions. The potential couplings will be instantiated depending on a specific context or situation later on.

The critical goal of a driver is to drive without any troubles, and it is represented as a generic function: To drive. This function generates an actual driving context or situation as its output, and both humans and machines initially detect it. This triggers the establishment of a driver's SA, and additional functions subsequently represent this SA process.

The initial detection process of SA is represented with the following three additional functions: To pay attention to traffic conditions; To pay attention to in-car display; To sense driving information by the system. The input of these functions are potentially coupled with the output of To drive. Moreover, the first two functions with their couplings are related to the establishment of the SA Lv. 1.

The establishment of the SA Lv. 1 leads to the development of the SA Lv. 2. Thus, an additional function: To comprehend current driving situation is defined. Also, this function is coupled with the upstream functions of To pay attention to traffic conditions and To pay attention to in-car display.

Further, the SA Lv. 3 is developed based on the establishment of the SA Lv. 2, which is represented by the additional function of To plan and identify next actions. The input of this function is connected with the output of To comprehend current driving situation, and the output provides the driver's decision whether to continue the autonomous driving or not. In the end, the decision provides the inputs of additional two functions: To continue autonomous driving and To take evasive actions, whose outputs are fed back to the input of the generic function: To drive.

Besides the above structure of this model, the output of To sense driving information is also coupled with the input of To alert drivers to prepare for emergency. This makes the downstream function ready for launching alerts. Moreover, the output of the alerting function is potentially coupled with the inputs of the functions (the functions of 2, 5, 6, and 9) related to each level of the driver's SA.

Table 6.2: Description about each FRAM function shown in Fig. 6.1.

1. To drive	
Input	Signals to continue autonomous driving
	Evasive actions
Output	Driving context
Precondition	Not Applicable (N/A)
Resource	N/A
Control	N/A
Time	N/A

2. To plan and identify next actions	
Input	Comprehension of current driving situation
	Alarms to instruct next actions
Output	Next actions
Precondition	N/A
Resource	N/A
Control	N/A
Time	N/A

3. To continue autonomous driving	
Input	Next actions
Output	Signals to continue autonomous driving
Precondition	N/A
Resource	N/A
Control	Driving context
Time	N/A

4. To take evasive actions	
Input	Next actions
Output	Evasive actions
Precondition	N/A
Resource	N/A
Control	Driving context
Time	N/A

5. To pay attention to traffic conditions	
Input	Driving context
	Alarms to warn something is wrong
Output	Observed traffic conditions
Precondition	N/A
Resource	N/A
Control	Observed information of autonomous driving
Time	N/A

6. To pay attention to in-car display	
Input	Driving context
	Alarms to warn something is wrong
Output	Observed information of autonomous driving
Precondition	N/A
Resource	N/A
Control	Observed traffic conditions
Time	N/A

7. To sense driving information	
Input	Driving context
Output	Sensed driving information
Precondition	N/A
Resource	N/A
Control	N/A
Time	N/A

8. To alert drivers to prepare for emergency	
Input	Sensed driving information
Output	Alarms to support the SA Lv. 1, 2, and 3
Precondition	N/A
Resource	N/A
Control	N/A
Time	N/A

9. To comprehend current driving situation	
Input	Observed traffic conditions
	Observed information of autonomous driving
	Alarms to tell what is going on
Output	Comprehension of current driving situation
Precondition	N/A
Resource	N/A
Control	N/A
Time	N/A

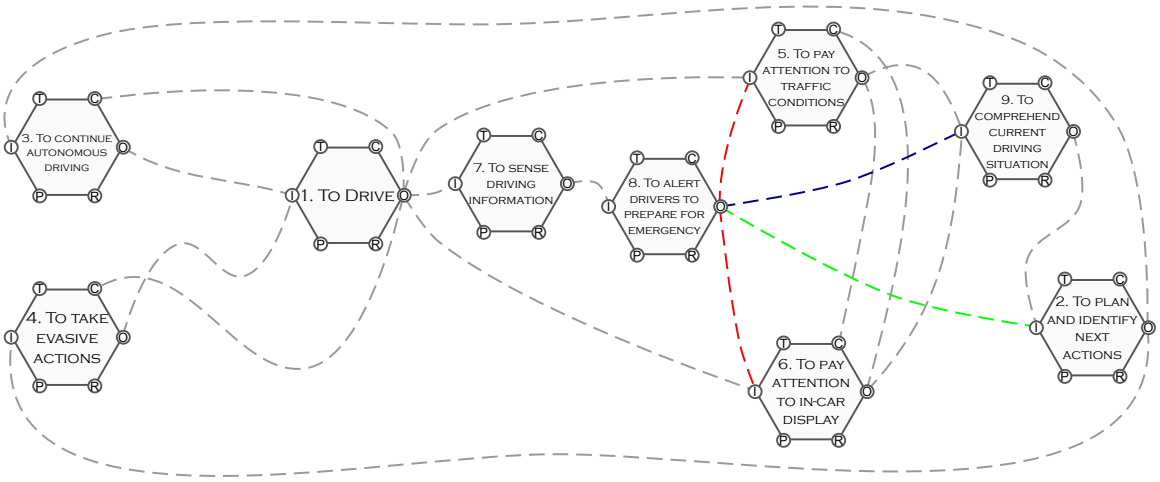


Fig. 6.1: Potential couplings among FRAM functions: operation of ADS.

By the way, the instantiation of potential couplings shown in Fig. 6.1 generates a large number of structures of FRAM model which is called instances; the number of possible instances indeed amount for $2^{19} = 524,288$ patterns in this case since there are 19 dotted lines, i.e., potential couplings in the general structure of FRAM model. However, it is apparently irrational to consider all of them, and the possible instantiations should therefore be limited on the basis of some constraints in advance. In this respect, the general structure of FRAM model in Fig. 6.1 can be reconfigured based on the perceptual cycle (Neisser, 1976; Smith and Hancock, 1995), and the constraints on the possible instantiations can be obtained based on its analysis.

The perceptual cycle describes a cyclic process in which cognitive agents (e.g., human beings or automation) take some actions directed by their internally held knowledge — or schemata of the world, and their outcome modifies the original knowledge or schemata as shown in Fig. 6.2. Specifically, the agents have a schema of present environment as a part of cognitive map of the world, and it directs a part of locomotion and action for perceptual exploration of the world. This exploration samples available information of actual present environment, and the sampled information updates the original schema of present environment. This cyclic relationship is invariant and must exist for any reasons. Based on this framework, the FRAM model in Fig. 6.1 can be redrawn as shown in Fig. 6.3.

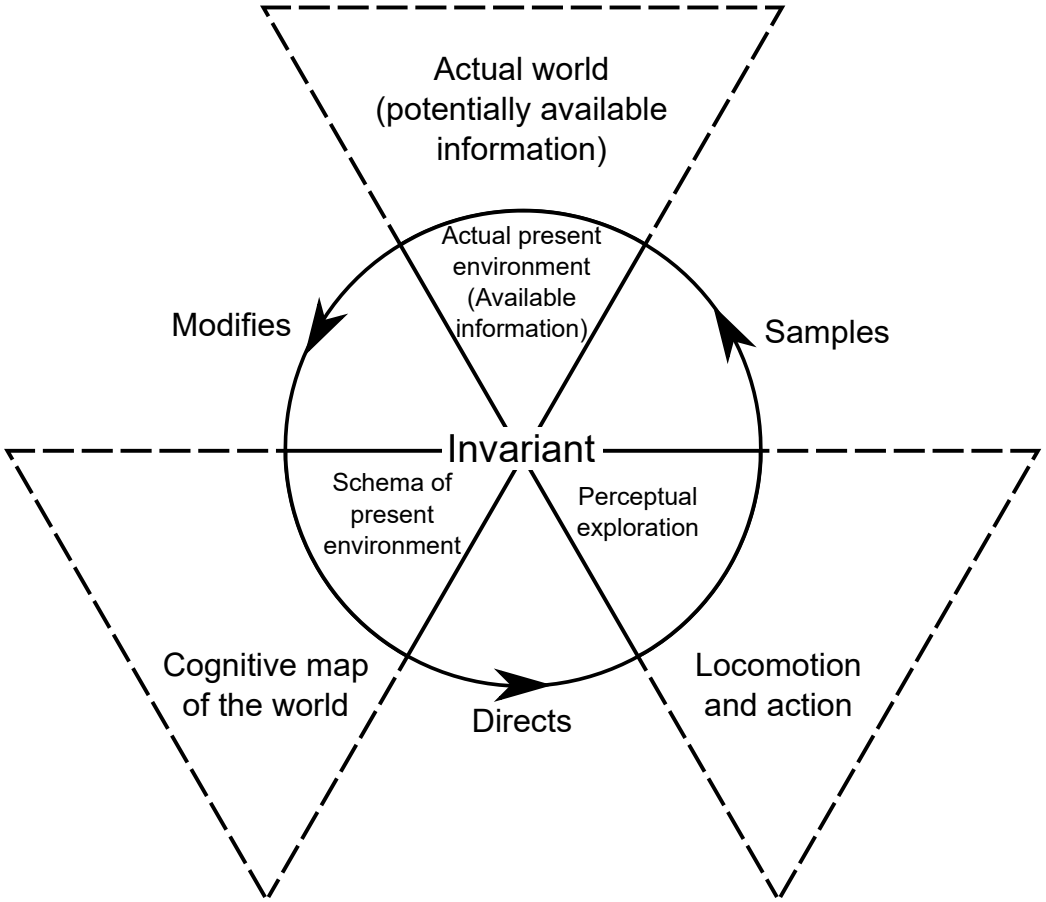


Fig. 6.2: Perceptual cycle (Neisser, 1976; Smith and Hancock, 1995).

Figure 6.3 illustrates another configuration of Fig. 6.1. In this configuration, each function is distributed and grouped into one of three elements of the perceptual cycle: To continue autonomous driving and To take evasive actions were grouped into the "Locomotion and action" of perceptual cycle; To drive was distributed to the "Actual world (potentially available information)" for convenience because the output of this function was originally defined to provide actual driving context or situation, rather than some specific actions as its output; the rest of functions were grouped into cognitive part, i.e., "Cognitive map of the world" of the perceptual cycle. Here, only the position of each function was changed, and other parameters such as potential couplings among functions are not modified at all. In addition, the grouped functions with potential couplings, i.e., dotted-lines are currently corresponding to the entire part of each triangle in Fig. 6.3; the functions get to represent the part of triangles clipped by the circle, i.e., the area with solid lines in Fig. 6.3 as a result of instantiations.

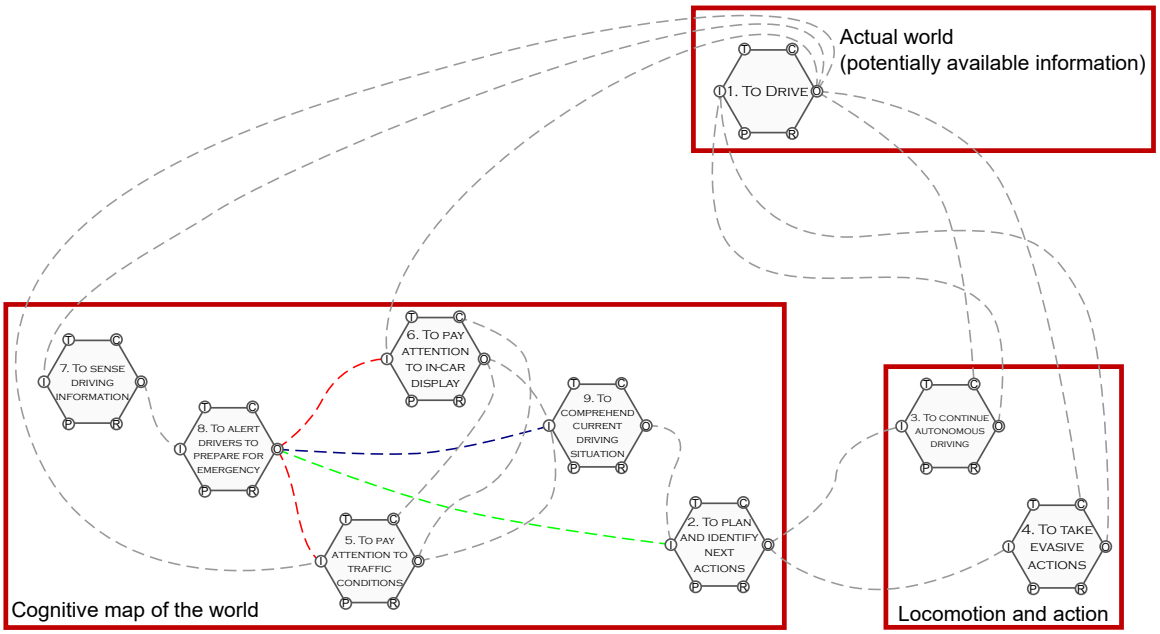


Fig. 6.3: Another configuration of FRAM model based on perceptual cycle.

A topological analysis of this reconfigured model provides at least three constraints on the instantiation. The first constraint is that dependency of functions interconnecting the groups, i.e., "Locomotion and action," "Actual world," and "Cognitive map of the world" of perceptual cycle must exist due to the invariance of their relationships. In other words, the potential couplings or functions forming this dependency basically have to be instantiated. The second constraint is that the two functions in the "Locomotion and action" group cannot be instantiated at the same time; if the function: To continue autonomous driving is instantiated and active, the other function: To take evasive actions should be deactivated. It is therefore allowed us to instantiate only potential couplings or functions belonging to the "knowledge" group with some freedom. In the end, the third constraint is that functions or potential couplings that we can arbitrary instantiate is limited to To pay attention to traffic conditions, To pay attention to in-car display, or the outputs of To alert drivers to prepare for emergency. This is because the rest of functions in this group are related to internal process of humans or machines, and their intentional instantiation could result in too much subjectivity/arbitrariness of the simulation. These constraints significantly limit the number of possible instantiations, and their actual patterns are provided in the following part of simulation scenario construction.

6.3.2 Parameters of CPCs

The original eleven CPCs shown in Table 3.2 is adopted in this simulation. Also, the CPC weight in each function is identified by using the AHP (Saaty, 1990) as well as the previous case study of steel production lines. The CPC weight in each function is consequently identified as shown in Table 6.3, where the identification numbers 1 – 11 in the first column correspond to that of CPCs shown in Table 3.2, and the identification numbers 1 – 9 after the second row correspond to that of the functions shown in Fig. 6.1 or Table 6.2.

Table 6.3: CPC weight in each function

ID No. of CPCs in Table 3.2	ID No. of functions shown in Fig. 6.1								
	1	2	3	4	5	6	7	8	9
1	0.102	0.240	0.0746	0.204	0.138	0.179	0.173	0.191	0.1600
2	0.102	0.0932	0.0746	0.0889	0.138	0.0640	0.0324	0.0213	0.0573
3	0.0355	0.0333	0.0284	0.0449	0.0462	0.0234	0.0324	0.0638	0.0573
4	0.102	0.0333	0.0746	0.0188	0.138	0.0640	0.0324	0.0638	0.1600
5	0.0355	0.0333	0.0284	0.0188	0.0154	0.0234	0.0324	0.0213	0.0214
6	0.102	0.0932	0.200	0.204	0.138	0.179	0.214	0.191	0.1600
7	0.102	0.0932	0.200	0.0889	0.138	0.179	0.173	0.191	0.1600
8	0.269	0.240	0.200	0.204	0.138	0.179	0.214	0.191	0.1600
9	0.102	0.0932	0.0746	0.0449	0.0462	0.0640	0.0324	0.0213	0.0214
10	0.0355	0.0333	0.0284	0.0628	0.0462	0.0234	0.0324	0.0213	0.0214
11	0.0151	0.0149	0.0153	0.0188	0.0154	0.0213	0.0324	0.0213	0.0191

6.3.3 Simulation Scenario

Simulation scenarios are prepared to envision the emerging functional behaviors, or how an event of the authority transfer may affect the driver's following cognitive activities exhaustively. The scenarios consist of the corresponding settings of multiple parameters such as CPC scores and potential couplings (i.e., instantiations) intervening into on-going simulation processes. The manual settings of these parameters can be regarded as additional variabilities, and they were classified into two categories in this simulation: the one is called exogenous variabilities causing the time-critical situations, triggered by the manual change of CPC scores; the other is called endogenous variabilities that are structural changes, i.e., instantiations of FRAM model in response to the exogenous variabilities. The simulation scenarios were constructed based on these two kinds of variabilities so that the exogenous variabilities cause a time-critical situation and induce the endogenous variabilities. That is, the following scenarios are to investigate the composite effect of not only the exogenous variabilities caused by the change of CPC scores but also the endogenous variabilities that

reconstruct the FRAM model during the on-going simulation process.

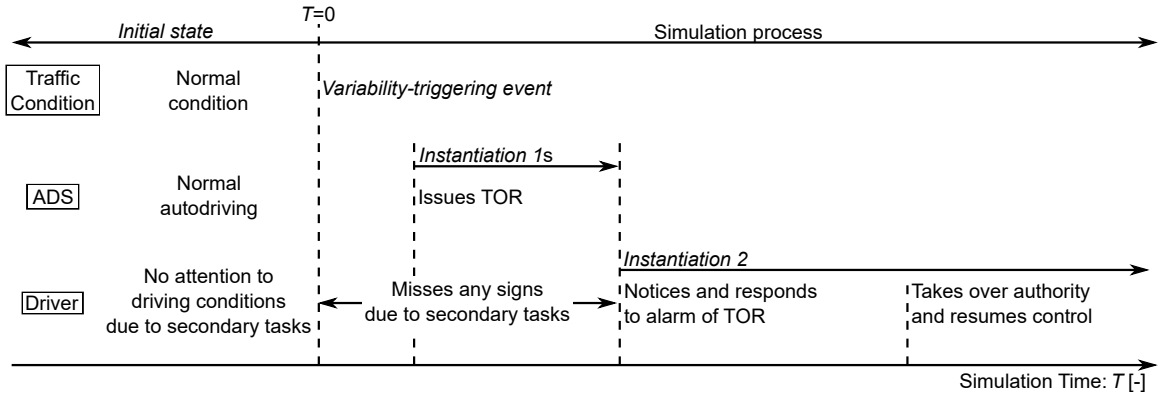


Fig. 6.4: Sequential diagram of *Verbal Description*.

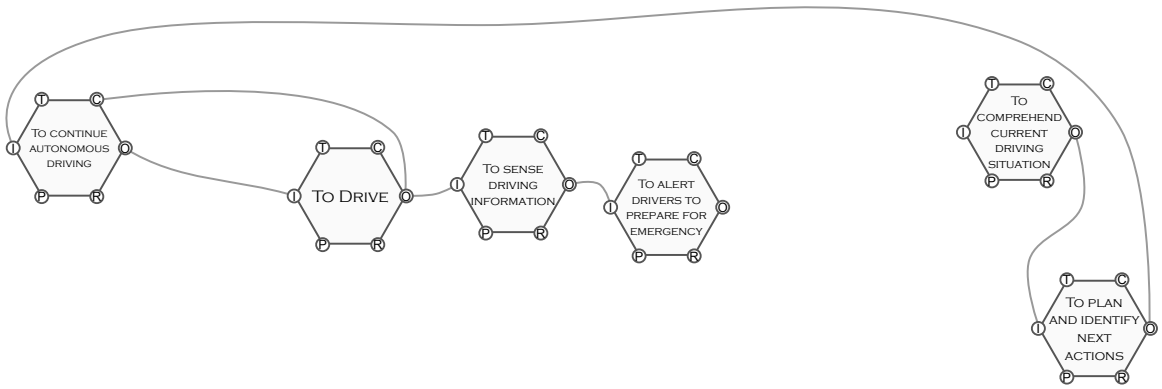


Fig. 6.5: A prototype of instantiations after the arrival of the alert: it will be common in the scenarios shown in Table 6.4.

The first step of the scenario construction is to prepare a verbal description of the scenario as follows:

> *Verbal Description*: A car is driving with its ADS, and the driver is currently distracted by some activities such as reading a book or using a smartphone rather than being engaged in the driving tasks, leaving them to the automation. Everything is going well at first. After a while, the system suddenly launches an alert demanding the driver to take over the authority of driving immediately. The driver must respond to this and take some evasive actions.

This verbal description can be represented as a sequential diagram in Fig. 6.4. The following steps of the scenario construction are to define several events described in

the diagram, i.e., *Initial state*, *Variability-triggering event* (exogenous variability), variations of *Instantiation 1*s, and *Instantiation 2* (endogenous variabilities), as the manual operation of simulation parameters.

First of all, the *Initial state* is set as following, representing the situation until the driver and ADS face a time-critical situation:

> *Initial state*: all CPC scores are equally set to 100, and the partial structure of Fig. 6.1 is instantiated, as shown in Fig. 6.5. Therein, two functions related to the SA Lv. 1 are temporally missing in the instance; this instance represents the active functional components existing during the autonomous driving without any involvement of the driver.

The simulation is then triggered by manually setting a CPC score as following at simulation time: $T = 0$. This manual operation is hereafter called *Variability-triggering event*, creating a time-critical situation to cause variabilities of the functions. Here, it should be noted that detailed descriptions of the situation such as disappearing lane markings, construction zones, or merging motorway lanes, all of which are popular factors to build scenarios of the authority transfer problem, are avoided to cope with the Envisioned World Problem.

> *Variability-triggering event*: the score of CPC: "Available time" is set to 0 at simulation time $T = 0$, and the simulation process is triggered.

After a while of the *Variability-triggering event*, the potential couplings inherent to the prototype shown in Fig. 6.5 are further instantiated. This intervention into the simulation process is to represent a situation in which the ADS launches an alert and request the driver to take over the driving authority immediately while the driver is still not aware of the emergency at all. This operation is hereafter referred to as *Instantiation 1*, for which three variations, i.e., *Instantiation 1a, b* and *c* are prepared; they are distinguished based on which level of the SA is supported by the warning system, and their details are shown in the following:

> *Instantiation 1a*: the potential couplings from To alert drivers to prepare for emergency to To pay attention to traffic conditions and To pay attention to in-car display are instantiated from the instance shown in Fig. 6.5. This represents the launch of alerts, forcing the missing two functions related to the SA Lv. 1 to activate.

Instantiation 1b: the potential coupling from To alert drivers to prepare for emergency to To comprehend current driving situation is instantiated. This represents the launch of alerts supporting the function related to the SA Lv. 2.
Instantiation 1c: the potential coupling from To alert drivers to prepare for emergency to To plan and identify next actions is instantiated. This represents the launch of alerts supporting the function related to the SA Lv. 3.

In the end, additional instantiation is carried out after any one of the variations in *Instantiation 1* in order to represent a situation in which the driver takes over the authority and resumes control. This intervention into the simulation process is hereafter referred to as *Instantiation 2*, and the detail is shown in the following:

Instantiation 2: To continue autonomous driving is deactivated, and To take evasive actions is activated; this represents the driver's responses to the issue of the alert, i.e., disengagement of the ADS and following evasive actions.

By temporally combining the above instantiations, we can build a sequence of instantiations in three different ways, each of which corresponds to the difference in how and when the system supports the driver's SA. This is shown in Table 6.4. Based upon these FRAM models corresponding to the three different scenarios, we simulate how the effect of the *Variability-triggering event* appears in each scenario and investigate into what the effective solutions for the authority transfer problem would be.

Table 6.4: Simulation scenarios.

Scenario No.	Flow of simulation scenario sequence
1	*Initial state* —> *Variability-triggering event* at $T = 0$ —> *Inst. 1a* —> *Inst. 2*
2	*Initial state* —> *Variability-triggering event* at $T = 0$ —> *Inst. 1b* —> *Inst. 2*
3	*Initial state* —> *Variability-triggering event* at $T = 0$ —> *Inst. 1c* —> *Inst. 2*

Inst.: *Instantiation*

Here, it is actually possible to consider additional situations and elaborate the scenario something like: "what if there would be additional alarms after the initial TOR" or "what if the driver finds another threat to avoid (e.g., obstacles) as a result of road inspection, and the time-critical situation occurs again during the evasive actions," for example. The former scenario can easily be realized by manipulating the

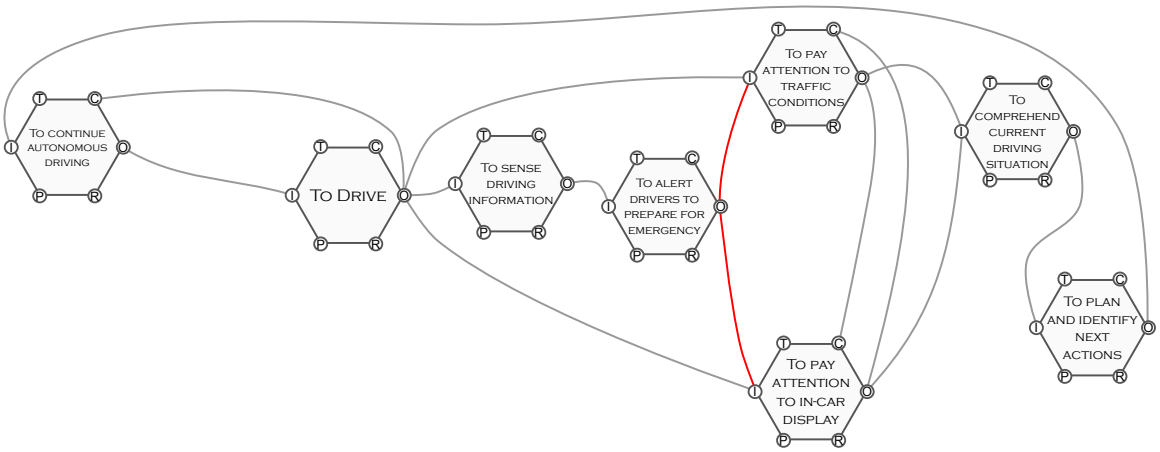
potential couplings coming from To alert drivers to prepare for emergency, and the latter scenario can be examined by setting the scores of CPCs: "Number of simultaneous goals" or "Available time" to 0, respectively after the *Instantiation 2*. Nevertheless, we tried simple cases in this simulation because the scenario construction should not take into account too many conditions or change of parameters at once.

One of the major limitations of this series of FRAM models is that the models essentially require complicated setting of parameters (e.g., functions, CPC weight, or simulation scenarios). In other words, too complicated parameter settings could easily lead to subjectivity or arbitrariness of simulations, and that is why the scenario should be constructed as simple as possible. Moreover, the focus of FRAM simulation is what kind of remarkable behavior can be observed based on the simple setting of parameters and their combinations, rather than collectively investigating complex cases; if there are still needs to consider more complicated scenarios, they should be extended based on the preceding — simpler simulations.

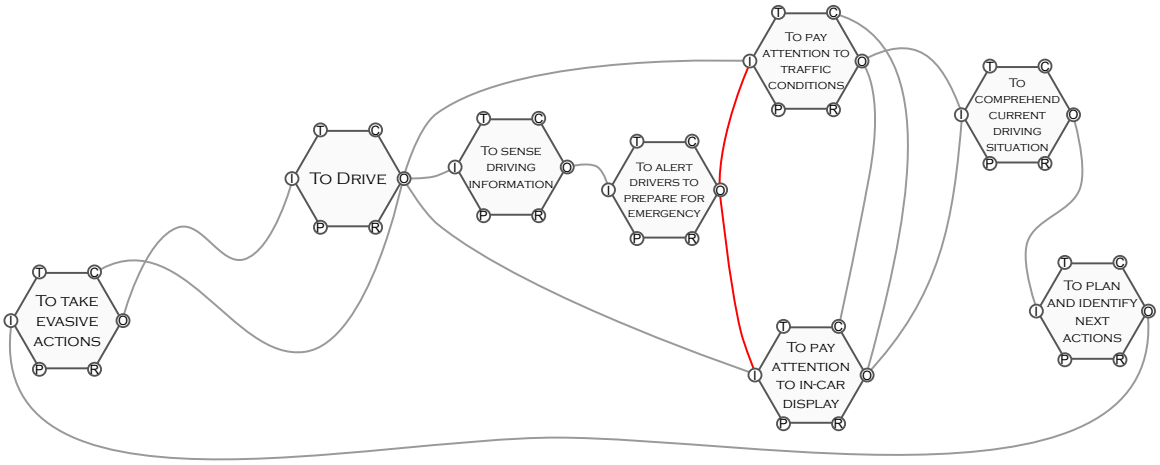
6.4 Simulations in Each Scenario and Result

6.4.1 Simulation Scenario 1

The first scenario is to investigate the effect of *Variability-triggering event* by the instantiations shown in Fig. 6.6. As stated in the previous subsection, the FRAM model of Fig.5(b) was built by chronically merging *Instantiation 1a* and *Instantiation 2* to original Instantiation 1. In the simulation, we varied the timings of *Instantiation 1a* and *Instantiation 2* in various ways and investigate how the effects of the *Variability-triggering event* would change accordingly. That is, the timing of the alert, and the timing of the driver's invoking SA Lv. 1. All combinations of the timings to cause the *Instantiation 1a* and *Instantiation 2* were simulated, and it was confirmed that the results could be classified into just three specific patterns shown in Fig. 6.7.



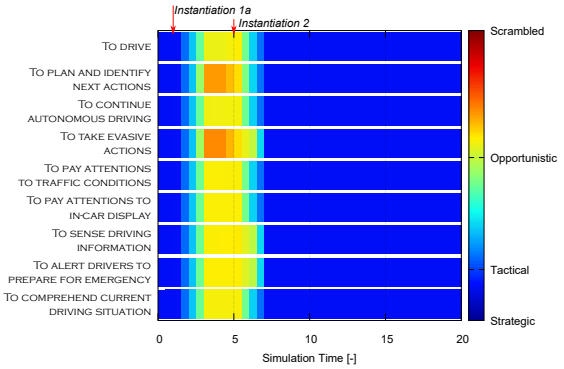
(a) Instance as a result of *Instantiation 1a*



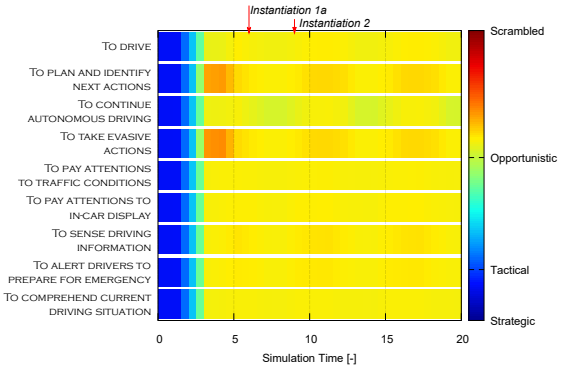
(b) Instance as a result of *Instantiation 1a* and *2*

Fig. 6.6: Instantiations in scenario 1.

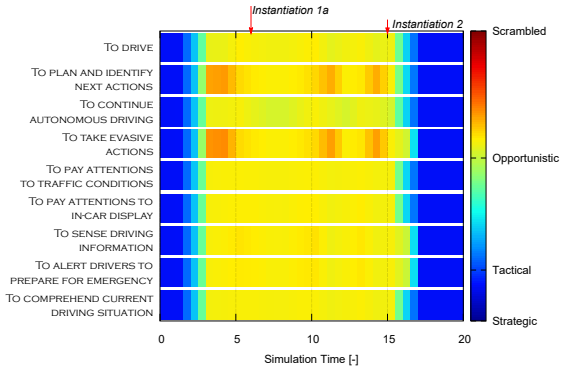
Figure 6.7(a) shows the transition pattern of the control modes in each function when *Instantiation 1a* and *Instantiation 2* are activated at simulation time $T = 1$ and $T = 5$, respectively. Wherein, the horizontal axis represents the simulation time, and color gradations represent the control mode in each function. As this pattern shows, the successive instantiations result in the outcome that the driver can make an immediate recovery of all functions from the effect of the *Variability-triggering event*. This recovery trend can be commonly observed when both the *Instantiation 1a* and *Instantiation 2* are activated earlier than the end of instantaneous peaks of control mode in To plan and identify next actions and To take evasive actions, existing around the simulation time $T = 3$–5. That is, the driver can recover from Variability, when both of the activations occur earlier than the simulation time T = 5 in this simulation.



(a) Pattern 1: *Instantiation 1a* at $T = 1$ and *Instantiation 2* at $T = 5$.



(b) Pattern 2: *Instantiation 1a* at $T = 6$ and *Instantiation 2* at $T = 9$.



(c) Pattern 3: *Instantiation 1a* at $T = 6$ and *Instantiation 2* at $T = 15$.

Fig. 6.7: Result of simulation scenario 1.

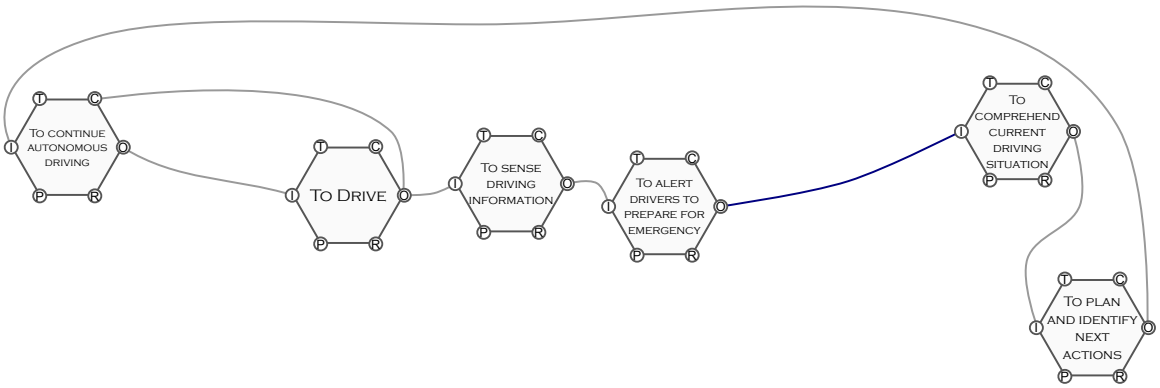
Figure 6.7(b) shows the next emerging pattern of transition. The difference from the setting of Fig. 6.7(a) is the delay of instantiating *Instantiation 1a* and consequent *Instantiation 2*. That is, it was delayed to inform the driver of the authority transfer (i.e., Take Over Request (TOR)), by which the driver can invoke his/her SA Lv. 1, and accordingly, the driver's taking evasive actions is also delayed. The result shows that it takes a longer time to recover from the *Variability-triggering event* than in the previous pattern. Moreover, the impact of the *Variability-triggering event* does never come to the stable within the time sequence of Fig. 6.7; an additional simulation confirms that the effect lasts until around the simulation time $T = 80$. This result shows that the delayed notification of the alert may cause the driver challenging to recover from the impact of the *Variability-triggering event*.

In addition to the above, another transition pattern emerges when the timing of *Instantiation 2* is shifted to further later. Wherein, the timing of *Instantiation 1a* is set the same with the previous case. In this pattern, the effect of the *Variability-triggering event* can be calmed down after the *Instantiation 2* as shown in Fig 6.7(c).

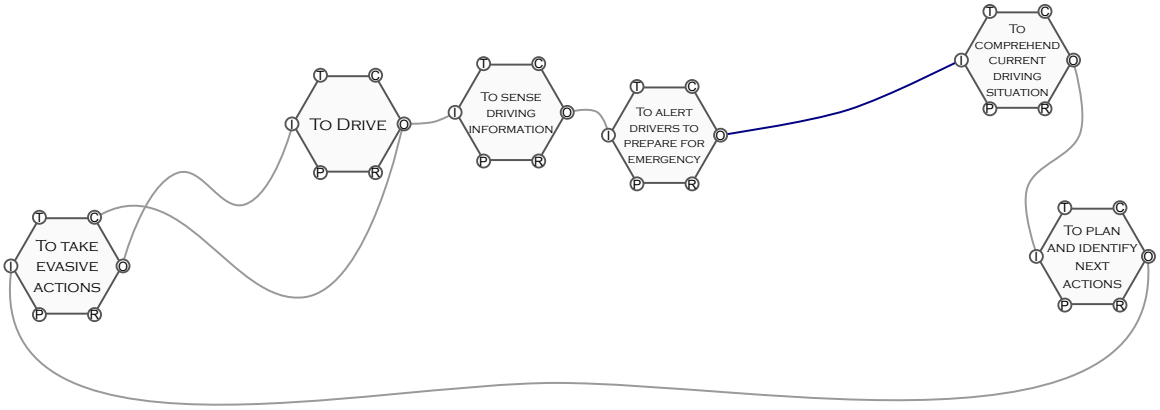
This trend can be observed even when the timing of *Instantiation 2* was delayed to the timing when the control modes of the functions of To plan and identify next actions and To take evasive actions are at the peaks (i.e., the most unstable), which exists around simulation time $T = 10$–15. This suggests that the countermeasures of being notified to invoke the driver's SA Lv.1 (i.e., *Instantiation 2*) would be useful if it would be taken earlier than the driver's confusion comes to be maximum, even though the notification of the TOR (i.e., Instantiation 1a) was delayed. This is further discussed later.

6.4.2 Simulation Scenario 2

The second scenario is to investigate the effect of the *Variability-triggering event* when the quality of support given by the system is changed from the previous case. The FRAM model used for this simulation is shown in Fig. 6.8. In this scenario, the role of the alarms is different from the previous scenario. It is supposed to provide some information telling the driver what is wrong and why, supporting the driver's SA Lv. 2. The driver needs to take some evasive actions based on it. The launch of the alarms is modeled as synchronized with the timings of *Instantiation 1b* and subsequent *Instantiation 2*. Wherein, *Instantiation 2* is the same as assumed in the previous Scenario 1.

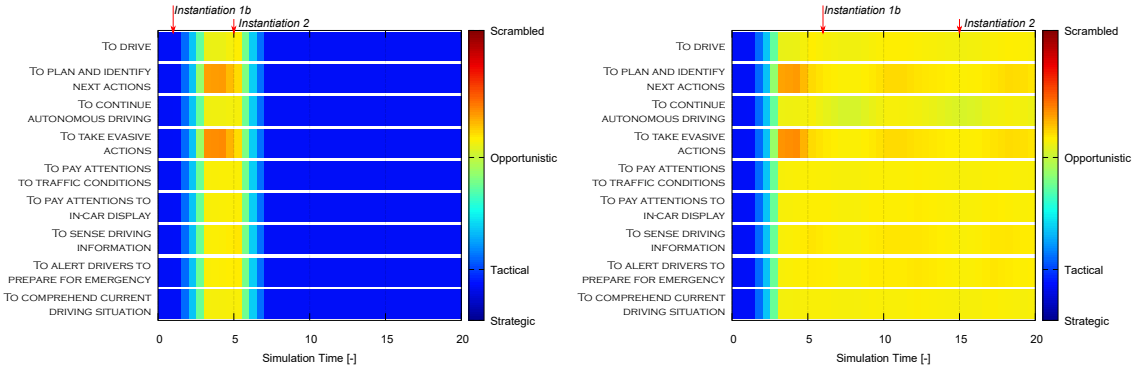


(a) Instance as a result of *Instantiation 1b*



(b) Instance as a result of *Instantiation 1b* and *2*

Fig. 6.8: Instantiations in scenario 2.



(a) Pattern 1: *Instantiation 1b* at $T = 1$ and *Instantiation 2* at $T = 5$.

(b) Pattern 2: *Instantiation 1b* at $T = 6$ and *Instantiation 2* at $T = 15$.

Fig. 6.9: Result of simulation scenario 2.

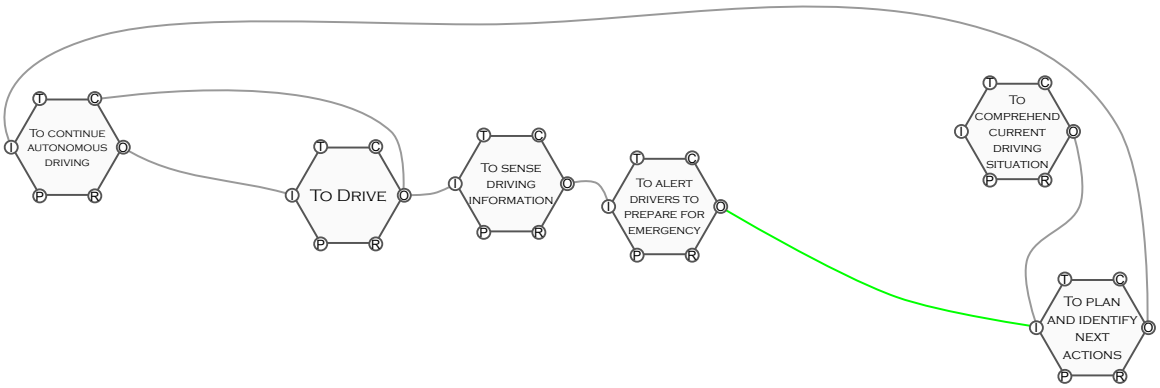
Figure 6.9(a) shows the case when *Instantiation 1b* and *Instantiation 2* are caused at simulation time $T = 1$ and $T = 5$, respectively. The timing of the instantiations is precisely the same as the case shown in Fig. 6.7(a). Also, the result is qualitatively

the same, suggesting that the countermeasures referred to as the *Instantiation 1b* and *Instantiation 2* can work effectively when they are taken in the early phase of the sequence. This trend lasts in the same way as the pattern 1 shown in Fig. 6.7(a). However, the emerging behaviors become quite different from the previous scenario shown in Fig. 6.7 when the alarms for the SA Lv. 2 come in delay.

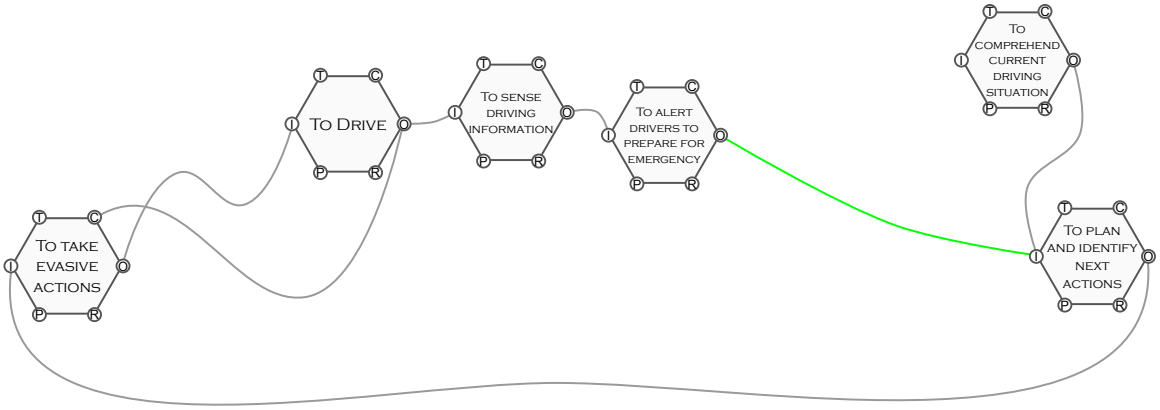
Figure 6.9(b) illustrates another transition pattern emerging in this scenario. Wherein, the timings of *Instantiation 1b* and *Instantiation 2* are set the same with the case of Fig. 6.7(c). However, the result is qualitatively different from the previous scenario drastically. The effect of the *Variability-triggering event* does never come to rest forever, even after the countermeasures are taken. This trend is similarly observed throughout the other combinations of timings to instantiate. All the results obtained in this scenario can eventually be classified into only two patterns shown in Fig. 6.9. According to the result, there exist chances to recover from the effect of the *Variability-triggering event* only when the instantiations were done in the early phase described in Fig. 6.7(a). Besides, any chances of recovery would be lost if they were done in delay. This suggests that the direct support of the SA Lv. 2 by the automation might be less effective than the support of the SA Lv. 1 in too time-critical situations.

6.4.3 Simulation Scenario 3

Figure 6.10 shows the instantiations in this scenario. The role of alarms in this scenario is to provide some information instructing the driver about actual evasive actions, supporting the SA Lv. 3 and their DM. The drivers are supposed to follow them. The launch of the alarms is represented as the *Instantiation 1c* shown in Fig. 6.10(a), and the evasive action is also instantiated as shown in the figure.



(a) Instance as a result of *Instantiation 1c*



(b) Instance as a result of *Instantiation 1c* and *2*

Fig. 6.10: Instantiations in scenario 3.

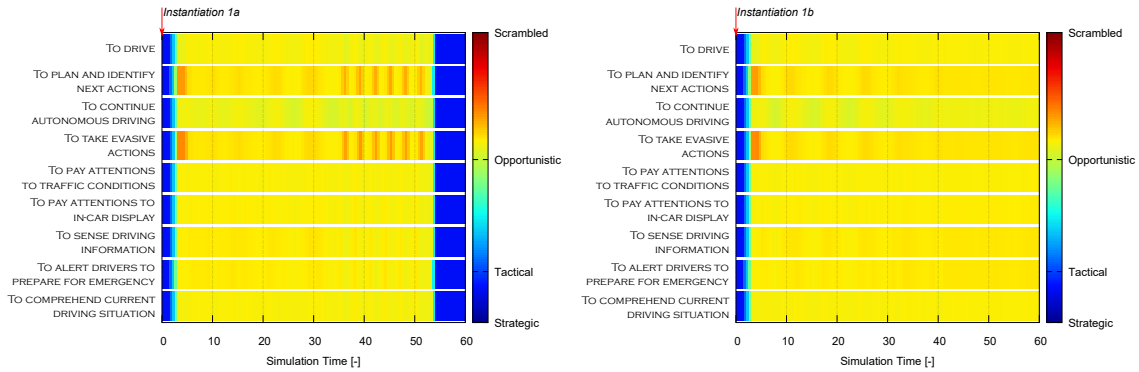
Figure 6.10 shows the instantiations in this scenario. The role of alarms in this scenario is to provide some information instructing the driver about actual evasive actions, supporting the SA Lv. 3 and consequent DM of evasive actions. The launch of the alarms is represented as the *Instantiation 1c* shown in Fig. 6.10(a), and the subsequent evasive action is modeled by *Instantiation 2* as shown in Fig. 6.10(b). All combinations of timing to cause the *Instantiation 1c* and *Instantiation 2* are simulated just like the other two scenarios. It was confirmed that the trend of the emerging transition patterns in this scenario are qualitatively the same as ones observed in the previous scenario of Fig. 6.9. That is, there would exist chances to recover from the effect of the *Variability-triggering event* only when the alarms come earlier, but the chances of the recovery were lost entirely otherwise. This also suggests that the direct support of the SA Lv. 3 by the automation might be less effective in too time-critical situations as well.

6.4.4 Relationships Between Simulation Results and Variation in *Instantiation 1*

According to the previous subsection, the simulation result in each scenario can be classified into two categories: the one is the simulation scenario 1 in which the countermeasure succeeds in the recovery from the effect of *Variability-triggering event*; the other is the simulation scenario 2 and 3 in which any chances of the recovery would be lost if the countermeasures are taken in delay. Now that the crucial difference among the simulation scenarios can be seen in the variation of the *Instantiation 1*, the variation seems to be related to the difference of the result. An additional simulation was conducted to verify this, and the effect of *Instantiation 1a*, *1b*, and *1c* is investigated here.

This simulation is, more specifically, to examine characteristics of the instances shown in Fig. 6.6(a), 6.8(a), and 6.10(a) in response to the effect of the *Variability-triggering event*. In this simulation, each of *Instantiation 1a*, *1b*, or *1c* was carried out first of all at the simulation time $T = 0$, and then the *Variability-triggering event* was also caused at $T = 0$. There were no interventions into the simulation process (e.g., *Instantiation 2*) after that, and the transition of the control mode was simulated until $T = 60$.

Figure 6.11(a) and 6.11(b) show the result when the *Instantiation 1a* and *1b* are carried out at $T = 0$, respectively; it should be noted that the result in the case of the *Instantiation 1c* is qualitatively the same as shown in Fig. 6.11(b), so we omit this case. Now we can confirm in Fig. 6.11(a) that the effect of the *Variability-triggering event* is calmed down in the case of the *Instantiation 1a* but keeps on remaining in the case of the *Instantiation 1b*; further simulation confirmed that the effect of *Variability-triggering event* lasts until around the simulation time $T = 700$ in the case of the *Instantiation 1b*.



(a) Response of the instance shown in Fig. 6.6(a).

(b) Response of the instance shown in Fig. 6.8(a).

Fig. 6.11: Responses of the instances shown in Fig. 6.6(a) and 6.8(a) against the effect of the *Variability-triggering event*.

This result can be explained from the perspective of static and dynamic stability of the FRAM model, which is usually discussed in aerodynamics of flight (FAA, 2016). The static stability in this context refers to a potential characteristic of a system against disturbances, depending on its inherent component such as structure or design; the static stability is positive when a system returns to the original state of equilibrium after being disturbed, neutral when a system remains in a new condition after its equilibrium has been disturbed, and negative when a system continues away from the original state of equilibrium after being disturbed. On the other hand, the dynamic stability refers to a time-dependent response of a system based on its static stability, disturbances, and additional input signals; the dynamic stability is positive when a displaced system decreases its amplitude of the displacement and returns toward an equilibrium state over time, neutral when the displacement neither decreases nor increases in amplitude, and negative when the amplitude of the displacement increases and becomes more divergent. Therefore, the result found that the static stability of the instance shown in Fig. 6.6(a) is positive, and that of the instance shown in Fig. 6.8(a) is substantially neutral since its recovery time is more than ten times longer than the case of the *Instantiation 1a*.

The simulation result further suggests the role of each instantiation: the variation of *Instantiation 1*s play a role in transforming the static stability of the FRAM model against the effect of the *Variability-triggering event*; *Instantiation 2* has effect on the dynamic stability of the FRAM model — or how fast the effect *Variability-triggering event* is calmed down. The dynamic stability generally exists only in the condition of the positive static stability, and the static stability was changed from neutral to positive only in the case of *Instantiation 1a*. That is why the composite effect of

Instantiation 1a and *2* succeeded in suppressing the effect of *Variability-triggering event*, and the effect of *Instantiation 1b* or *1c* and *2* failed in the late phase of the simulation sequence. Therefore, from the perspective of the functional abstraction level, how to establish the positive static stability of the FRAM model is the highest priority, and how and when to respond to TOR should be of secondary importance.

6.4.5 Brief Summary of Simulations

We confirmed that all FRAM models for the three simulation scenarios can work well and are expected to contribute to verifying how an event of the authority transfer affects the performance of the driving and what the restrictions for those. Furthermore, a countermeasure to recover from performative variabilities caused by the authority transfer was also validated. This was consistent with our intuition because we have been simulating the authority transfer in a time-critical situation, and the result suggests that the faster is better. Also, the result gained from the FRAM model of Fig. 6.6 verified that the system's notification at the level of a driver's SA Lv. 1 turned out to be effective even after the instruction of TOR to a driver is delayed, and the recovery mechanism is related to the static stability of the FRAM model at the functional abstraction level.

This is crucial because the reaction time of human drivers is generally limited, and it is less likely for the drivers to react immediately to the abrupt instruction. Notably, the reaction time could be even longer if they were occupied with some other secondary tasks.

This leads us to conclude that the driver's SA should be left in the control loop, and the system's support for the driver's SA should be designed so that the drivers could be aware of the situation from Lv. 1 by themselves. On the other hand, if the system's support for the driver's SA starts from Lv. 2 and Lv. 3, the confusion of the driver might be more, and they could not be adequate to let them recover from the chaos. The possible reason for this is that the information related to the SA Lv. 2 and 3 is generally so ad-hoc that it does not support their contextual awareness and strategic actions. This results in the driver's opportunistic reactions, and any chances of the recovery are lost in the end. Therefore, the establishment of SA Lv. 1 could play a significant role in conditional driving automation.

6.5 Discussions: Feasibility of Conditional Driving Automation

The simulation result suggests that the drivers should pay attentions to the driving conditions even if they are allowed to be free from the Dynamic Driving Tasks (DDTs) by the conditional driving automation or LoDA 3. However, the history of automation has been showing that it is intrinsically very burdensome for the human beings to keep on monitoring the systems and maintain their SA. This human characteristic has been proved by the professionals of the automation such as airline pilots or operators of nuclear power plants, and it is not difficult to imagine that the situation could be even worse in the case of non-professionals of the automation (i.e., usual human drivers). Therefore, it is now questioning whether the conditional driving automation is feasible in the real field of practice. Edwards et al. (2016), Dehais et al. (2012), and Inagaki and Sheridan (2019) discussed the important points related to this problem.

Edwards et al. (2016) have experimented with investigating the relationship between workload modified by the automation support, SA, and responding time, i.e., performance of air traffic controllers. They found through their experiment that the too much support of the automation could result in a reduction of engagement, leading to reduced SA and performance. They eventually concluded that the automation should not be designed just to reduce the workload; the automation should be designed to support the human operators to maintain their contextual SA.

In terms of the contextual SA, Dehais et al. (2012) have also conducted an experiment to investigate how conflicts of intentions between human operators and automated systems affect the human operators' performance. The result has confirmed that the conflicts occurred during the experimental operation of an Unmanned Ground Vehicle (UGV) led to perseveration behavior, higher heart rate, excessive-limited attentional focus of a major part — 69.2% of the participants, having resulted in the failure of their given task. They pointed out based on this result that automation overriding human operators' actions such as protection systems could cause this kind of conflicts and perseveration behaviors. Moreover, such operational supports or interventions by the automation alone are meaningless as long as the human operators are out of the loop and does not understand their behavior. The priority of automation design should, therefore, be put more on how to "cure" persevering human operators when the operators face a conflict, for which one possible solution is to identify and control the conflicts so that the human operators reconstruct their persevering SA.

Inagaki and Sheridan (2019) have pointed out the possibility that the drivers could fail to respond to a TOR during the autonomous driving at LoDA 3, for which additional operational supports such as steering or applying brake by the automation might be necessary. They proved this with a mathematical evaluation, suggesting that the smooth authority transfer cannot be achieved without some highly automated operational supports. On the other hand, they also discussed that this evaluation is no longer consistent with the framework of the LoDA 3, which cannot be classified into the single category; it should be noted that the evaluation also conflicts with the suggestions provided by Dehais et al. (2012). In the end, they concluded that the LoDA 3 is not the level where the R&D of the autonomous driving should aim and instead proposed two solutions: the one is to revise the definition of the conditional driving automation issued by SAE (2016); the other is to design HMI enhancing the human-machine collaboration so that the driver can smoothly take over the authority.

Our conclusion is consistent with theirs: the drivers must still maintain good SA during the autonomous driving while the system should also issue the TOR with its highly automated operational supports, if necessary. This suggests that human-machine collaboration still plays a significant role there, and the nature of the authority transfer problem is ambiguous rather than explicit. Both humans and machines are always responsible for driving, and it is irrational to consider the role of them individually. The focus should, therefore, be put on how to realize the productive human-machine collaboration throughout the driving sequence.

As an example, Cropley (2019) has reported that speed limiters and driver monitors will be mandatory in EU from 2022 and pointed its negative aspects from the perspective of human cognitions, all of which are consistent with the discussions so far. No matter how the importance of productive human-machine collaboration has been stressed for decades, the reality is that people still want to entrust everything to technologies in this way. This kind of approach actually sounds reliable and socially acceptable at a glance since human beings are still regarded as something like "sources of errors." However, our simulation result suggests that this is no longer true. The productive human-machine collaboration must not be neglected, no matter how sophisticated automation supports are introduced.

One possible solution for the productive human-machine collaboration is to develop an HMI or some other mechanisms with which the drivers can keep their attention to the driving. The simulation result has shown that the SA should be established without missing any levels. Moreover, the establishment of the SA Lv. 1 must be

supported by the system with the highest priority, and the more automated cognitive support for the SA Lv. 2 or 3 should follow it even if necessary; otherwise, the performance of the drivers could be reduced rather than enhanced. The support of the SA Lv. 1 can be achieved without any complicated, high technological equipment since the drivers only have to notice trivial elements related to the current situation. Specifically, they just need to know or feel "everything is going well" or "something is wrong" during the autonomous driving and the further SA should be developed by themselves.

In this respect, Vanderhaegen (2016, 2017) has addressed the related issue from the perspective of cognitive dissonance (Festinger, 1957). The cognitive dissonance is an incoherence among cognitions of human beings. The dissonance occurs when something seems to be wrong or different, and we feel gaps or conflicts among our individual or collective knowledge; it may generate our discomfort or a situation overload, and the discomfort may recursively result in further dissonance, whose process enables us to discover new knowledge or cognitions. In this context, Vanderhaegen (2016) proposed an original rule-based tool to model human and technical behaviors, detect possible conflicts among them, and assist the discovery and control of the dissonance; it should be noted that this approach can be associated with the identification and control of the conflicts suggested by Dehais et al. (2012). In addition, his succeeding study (Vanderhaegen, 2017) suggested that such unstable conditions, i.e., dissonance could contribute to adaptive behaviors of human beings or the resilience of systems.

An aspect of this approach can be confirmed in our result as well. The result shown in Fig. 6.7(c) suggests that the instantaneous peak of the control mode, i.e., unstable condition in two functions: To plan and identify next actions and To take evasive actions could contribute to the immediate recovery from the effect of variabilities. This result is therefore consistent with the concept of dissonance in terms of using unstable for stable/resilience. It is essential to take into account these aspects for the productive human-machine collaboration, and the future work will be conducted to elucidate its nature.

6.6 Brief Summary

To investigate the nature of the authority transfer at LoDA 3 in time-critical situations, a simulation based on FRAM was carried out. The result has shown that the driver's SA should be left in the control loop even during the autonomous driving.

Specifically, the system's support for the driver's SA should be designed so that the drivers could be aware of the situation from Lv. 1 by themselves, otherwise any chances of the recovery in the authority transfer sequence would be lost. This also corresponds to transforming the static stability of the FRAM model from neutral to positive at the functional abstraction level, and maintaining the positive static stability should be the highest priority. Therefore, the positive static stability plays a significant role in the authority transfer problem, and it is useless to try any measures without considering the static stability of the functional model; the transformation of the static stability is triggered by the establishment of SA Lv. 1, according to this simulation result.

While the simulation result suggests that the driver's SA should be established without missing any levels, the history of automation has been showing that it is intrinsically very burdensome for the human beings to keep on monitoring the systems and maintain their good SA. The focus should, therefore, be put on how to realize the productive human-machine collaboration throughout the driving sequence, and it is irrational to consider the role of them individually; the human must be more or less engaged in DDTs anytime as long as the full — 100% automation was not invented. This is also the case with other automation technologies, and therefore, the human-machine collaborations must not be neglected even if highly sophisticated automations are developed.

Chapter 7

Conclusion

This thesis has provided simulation models to envision emergent behavior, or safety of artifacts as socio-technical systems based on FRAM and addressed potential problems inherent to three principles to design human-machine systems.

The simulation models based on FRAM have been developed by integrating the concept of FRAM and a methodology of Fuzzy CREAM which is an extended model of the Cognitive Reliability Error Analysis Method (CREAM). The Fuzzy CREAM enabled us to introduce numerical definition of the variabilities existing in functions and their surrounding working environment. Also, the numerical definition of those variabilities made it possible to formulate interactions of the variabilities. The initial model was then updated twice based on the ideas of the Cellular Automaton and the Complex Adaptive Systems, respectively. Each of three model was adopted for case studies to investigate potential problems inherent to three principles to design human-machine systems , i.e., Compensatory Principle, Leftover Principle, and Complementarity/Congruence Principle.

The first case study was to investigate the validity of the Compensatory Principle in ever changing environment. The potential problem with this principle is that the roles of human and machines often tend to be fixed, and any flexibilities of those systems' operations could be lost in real fields of practice; it is therefore necessary for this principle to take into account the feasibility of such static function allocations in ever changing environment. In this respect, the initial FRAM model has been adopted for the analysis of an actual air crash accident, which is assumed to be triggered by complex design of Human-Machine Interface (HMI), SOPs, and subsequent loss of situation awareness of the pilots. It was shown in the analysis that how the feasibility of a SOP carried out in the accident sequence was changed by variabilities of communication quality, available time, and complex design of the HMI. Also, the analysis result found that there could have been a possibility to survive if the SOP

had properly been modified and carried out in the same situation. Therefore, the feasibility of SOPs plays a significant role in the safety of automated systems. The simulation consequently confirmed that the proposed FRAM model can address the potential problem of the Compensatory Principle as a stress test tool of artifacts, including not only physical objects but also intangible products such as manuals or checklists; the stress test can specifically contribute to the resilience engineering in terms of how to achieve the proper design SOPs and their flexible modifications.

The second case study was to address a problem inherent to the Leftover Principle. The characteristic of this principle is that while everything found to be feasible is automated, human beings are pushed to a domain where too complex tasks/activities to be automated are left and need to cope with all of them. Operations of systems based on this principle shall, consequently, depend on implicit knowledge and skills of human operators, and it is generally difficult to justify their validity. To address the issue, the second — Cellular Automaton based FRAM model has been applied to elucidate the nature of an empirical knowledge inherited in operations of steel production lines, i.e., adjustment of the rate of direct delivery. The simulation result found that the success of the adjustment is closely related to a timing when it is carried out; there were several timings when the target steel production line becomes "high-entropy" state, and the adjustment can effectively work when it is conducted at these points. The similar phenomena can be observed in behavior of Complex Adaptive Systems. This case study, in conclusion, provided an insight that it is essential for the safety management of socio-technical systems to consider, understand, and harness such complex dynamics of systems; the human beings in the Leftover Principle empirically cope with such complexities, and the fact has been envisioned as a result of simulation.

The third case study was to demonstrate the importance of the Complementarity/Congruence Principle, or more specifically, human-machine collaborations in the future. In the near future, our daily lives (e.g., houseworks or driving cars) will be automated with high technologies, for which safety effective human-machine collaborations play a significant role. However, this point of view often are often neglected since people tend to think the highly sophisticated automations are capable of everything, and human-machine collaborations are not necessarily important. To discuss this is not true, the safety and feasibility of the SAE conditional driving automation (LoDA 3) has been investigated with the third — Complex Adaptive Systems based FRAM model. The simulation result found that the involvement of human drivers

in driving activities is still essential even during autonomous driving, and the current LoDA 3 is a myth that the human drivers are supposed to be required just in non-normal situations, while they can be free from Dynamic Driving Tasks (DDTs) in usual cases. It is, therefore, much more important and effective to focus on how to realize the productive human-machine collaboration throughout the driving sequence, rather than considering their role individually like popular approaches.

The case studies so far have confirmed that each of them provided insights or criteria to address the safety issues from the functional perspective. Basically, these products are apparently beyond what can be obtained with traditional approaches based on reductionism (e.g., reliability engineering). The proposed FRAM models have surely extended the possibility of safety analysis in this respect; they eventually take on a role of the functional part in the envisioned world problem and provide new insights or criteria to facilitate the discussions of the functional safety as shown in Fig. 7.1. The right design, or Verification and Validation (V&V) of artifacts should be conducted from the multi-abstraction levels in this way, and its importance must increase in the future.

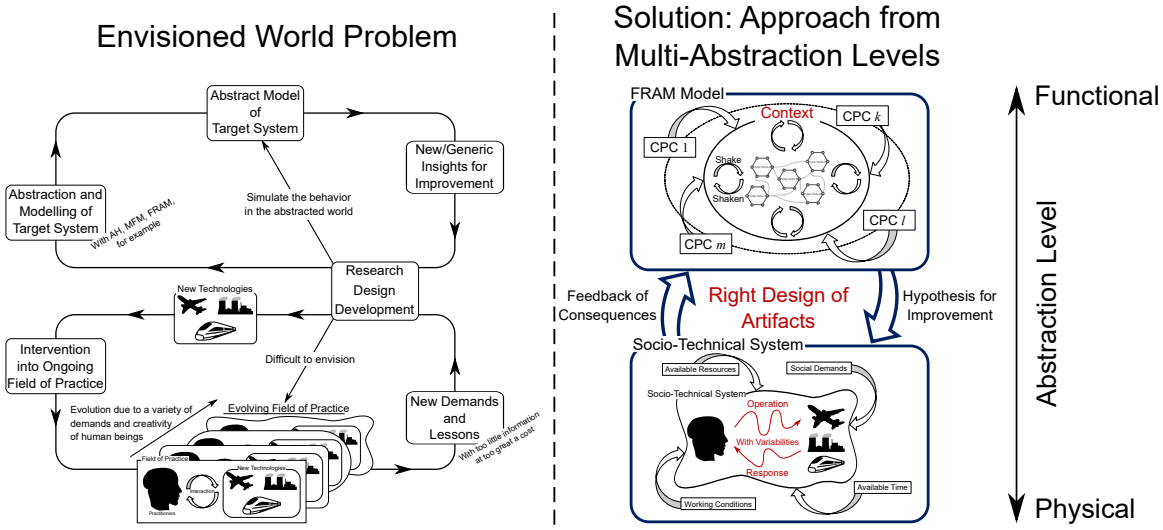


Fig. 7.1: Correspondence of FRAM to envisioned world problem: schematic illustration of right design of artifacts.

In the end, this research will go on to further step into the nature of the safety. One breakthrough I currently believe is that the functional resonance or complexity of socio-technical systems could provide not only negative but also positive consequences as suggested in the FRAM simulations. Its explorations shall lead us to a new dimension of the safety.

Bibliography

AIB, 2019. Aircraft Accident Investigation Preliminary Report Ethiopian Airlines Group B737-8 (MAX) Registered ET-AVJ 28 NM South East of Addis Ababa, Bole International Airport March 10, 2019. Accident Report AI-01/19. Federal Democratic Republic of Ethiopia Ministry of Transport Aircraft Accident Investigation Bureau.

Akyuz, E., Celik, M., 2015. Application of CREAM human reliability model to cargo loading process of LPG tankers. Journal of Loss Prevention in the Process Industries 34, 39–48.

Barriere, M., Bley, D., Cooper, S., Forester, J., Kolaczkowski, A., Luckas, W., Parry, G., Ramey-Smith, A., Thompson, C., Whitehead, D., et al., 2000. Technical basis and implementation guidelines for a technique for human event analysis (ATHEANA). NUREG-1624, Rev 1.

Berlekamp, E.R., Conway, J.H., Guy, R.K., 1982. Winning Ways for your Mathematical Plays, volume 2. AK Peters/CRC Press.

Bialek, W., Cavagna, A., Giardina, I., Mora, T., Pohl, O., Silvestri, E., Viale, M., Walczak, A.M., 2014. Social interactions dominate speed control in poising natural flocks near criticality. Proceedings of the National Academy of Sciences 111, 7212–7217.

Brauer, W., Reisig, W., 2009. Carl Adam Petri and "Petri Nets". Fundamental Concepts in Computer Science 3, 129–139.

Bytheway, C.W., 2007. FAST Creativity and Innovation: Rapidly Improving Processes, Product Development and Solving Complex Problems. J. Ross Publishing.

Cantelon, P.L., Williams, R.C., 1980. Crisis Contained: The Department of Energy at Three Mile Island. Southern Illinois University Press.

CAPA, 1996. AA965 Cali Accident Report: CONTROLLED FLIGHT INTO TERRAIN AMERICAN AIRLINES FLIGHT 965 BOEING 757-223, N651AA NEAR CALI, COLOMBIA DECEMBER 20, 1995 . Accident Report. AERONAUTICA CIVIL of THE REPUBLIC OF COLOMBIA.

Carpenter, S., Walker, B., Anderies, J.M., Abel, N., 2001. From Metaphor to Measurement: Resilience of What to What? Ecosystems 4, 765–781.

Carroll, J.M., Campbell, R.L., 1989. Artifacts as psychological theories: The case of human-computer interaction. Behaviour & Information Technology 8, 247–256.

Chiles, J.R., 2002. Inviting Disaster: Lessons From the Edge of Technology. HarperCollins.

Clarke, E.M., Grumberg, O., Browne, M.C., 1987. Reasoning about Networks with Many Identical Finite-State Processes. Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.

Cropley, S., 2019. Speed limiters may create more dangers than they prevent: EU's decision to make speed limiters mandatory on all new cars from 2022 plays to the view that we're overly reliant on car tech. `https://www.autocar.co.uk/opinion/industry/speed-limiters-may-create-more-dangers-they-prevent`. Retrieved on Aug. 4th 2020.

Degani, A., 2004. Taming HAL: DESIGNING INTERFACES BEYOND 2001. Springer.

Degani, A., Wiener, E.L., 1994. On the Design of Flight-Deck Procedures. NASA Contractor Report. NASA Ames Research Center.

Dehais, F., Causse, M., Vachon, F., Tremblay, S., 2012. Cognitive conflict in human–automation interactions: A psychophysiological study. Applied ergonomics 43, 588–595.

Duan, G., Tian, J., Wu, J., 2015. Extended FRAM by integrating with model checking to effectively explore hazard evolution. Mathematical Problems in Engineering 2015.

Edwards, T., Homola, J., Mercer, J., Claudatos, L., 2016. Multifactor interactions and the air traffic controller: The interaction of situation awareness and workload in association with automation. IFAC-PapersOnLine 49, 597–602.

Endsley, M.R., 1995a. Toward a Theory of Situation Awareness in Dynamic Systems. Human Factors The Journal of the Human Factors and Ergonomics Society 37, 32–64.

Endsley, M.R., 1995b. TOWARDS A NEW PARADIGM FOR AUTOMATION: DESIGNING FOR SITUATION AWARENESS. IFAC Proceedings Volumes 28, 365–370.

Endsley, M.R., 2001. Designing for Situation Awareness in Complex Systems, in: Proceedings of the Second International Workshop on symbiosis of humans, artifacts and environment, pp. 1–14.

Engeström, Y., 1987. Learning by Expanding: An Activity-theoretical Approach to Developmental Research. Orienta-Konsultit Oy.

Eriksson, A., Stanton, N.A., 2017. Takeover Time in Highly Automated Vehicles: Noncritical Transitions to and From Manual Control. Human Factors 59, 689–705.

FAA, 2016. Pilot's Handbook of Aeronautical Knowledge. Handbooks & Manuals FAA-H-8083-25B. Federal Aviation Administration.

Festinger, L., 1957. A Theory of Cognitive Dissonance. volume 2. Stanford university press.

Fitts, P.M., 1951. Human engineering for an effective air-navigation and traffic-control system. .

Fu, Z., Zhou, X., Zhu, K., Chen, Y., Zhuang, Y., Hu, Y., Yang, L., Chen, C., Li, J., 2015. A floor field cellular automaton for crowd evacuation considering different walking abilities. Physica A: Statistical Mechanics and its Applications 420, 294–303.

Fukui, M., Ishibashi, Y., 1996. Traffic flow in 1D cellular automaton model including cars moving with high speed. Journal of the Physical Society of Japan 65, 1868–1870.

George, J.K., Bo, Y., et al., 1995. Fuzzy sets and fuzzy logic: Theory and applications. PHI New Delhi , 443–455.

Heinrich, Herbert, W., 1931. Industrial accident prevention; a scientific approach. McGraw-Hill: New York.

Holling, C.S., 1973. RESILIENCE AND STABILITY OF ECOLOGICAL SYSTEMS. Annual review of ecology and systematics 4, 1–23.

Hollnagel, E., 1998. Cognitive reliability and error analysis method (CREAM). Elsevier.

Hollnagel, E., 2003. Context, cognition and control. Co-operative process management , 27–52.

Hollnagel, E., 2004. Barriers and Accident Prevention. Ashgate Publishing Ltd.

Hollnagel, E., 2012a. CREAM - Cognitive Reliability and Error Analysis Method. `http://erikhollnagel.com/ideas/cream.html`. Retrieved on Nov. 15th 2019.

Hollnagel, E., 2012b. FRAM: The Functional Resonance Analysis Method: Modelling Complex Socio-technical Systems. Ashgate Publishing Ltd.

Hollnagel, E., 2017. SAFETY-II IN PRACTICE: DEVELOPING THE RESILIENCE POTENTIALS. Taylor & Francis.

Hollnagel, E., Bye, A., 2000. Principles for modelling function allocation. International Journal of Human-Computer Studies 52, 253–265.

Hollnagel, E., Goteman, O., 2004. The Functional Resonance Accident Model. Proceedings of cognitive system engineering in process plant 2004, 155–161.

Hollnagel, E., Leonhardt, J., Licu, T., Shorrock, S., 2013. From Safety-I to Safety-II: A White Paper: Eurocontrol.

Hollnagel, E., Woods, D.D., 2006. JOINT COGNITIVE SYSTEMS: Patterns in Cognitive Systems Engineering .

Hollnagel, E., Woods, D.D., Leveson, N., 2006. Resilience Engineering: Concepts and Precepts. CRC Press.

Holzmann, G.J., 2004. The SPIN model checker: Primer and reference manual. volume 1003. Addison-Wesley Reading.

IAEA, 1999. Report on the Preliminary Fact Finding Mission Following the Accident at the Nuclear Fuel Processing Facility in Tokaimura, Japan. Accident Report. International Atomic Energy Agency.

IAEA, 2015. The Fukushima Daiichi Accident. Accident Report STI/PUB/1710. International Atomic Energy Agency.

ICAO, 2009. Safety Management Manual (SMM). Second ed., International Civil Aviation Organization.

Inagaki, T., 2003. Adaptive Automation: Sharing and Trading of Control. Handbook of Cognitive Task Design 8, 147–169.

Inagaki, T., Moray, N., Itoh, M., 1998. TRUST, SELF-CONFIDENCE AND AUTHORITY IN HUMAN-MACHINE SYSTEMS. IFAC Proceedings Volumes 31, 431–436.

Inagaki, T., Sheridan, T.B., 2019. A critique of the SAE conditional driving automation definition, and analyses of options for improvement. Cognition, Technology & Work 21, 569–578.

Israeli, N., Goldenfeld, N., 2004. Computational irreducibility and the predictability of complex physical systems. Physical review letters 92, 074105.

Johnson, N., 2009. SIMPLY COMPLEXITY: A Clear Guide to Complexity Theory. Oneworld Publications.

JTSB, 1996. AIRCRAFT ACCIDENT INVESTIGATION REPORT China Airlines Airbus Industrie A300B4-622R, B1816 Nagoya Airport April 26, 1994. Accident Report 96–5. Aircraft and Railway Accidents Investigation Commission.

JTSB, 2007. Train Derailment Accident between Tsukaguchi and Amagasaki Stations of the Fukuchiyama Line of the West Japan Railway Company. Accident Report RA2007-3-1. Aircraft and Railway Accidents Investigation Commission.

Kirlik, A., 1993. Modeling strategic behavior in human-automation interaction: Why an" aid" can (and should) go unused. Human factors 35, 221–242.

Kirwan, B., Ainsworth, L.K., 1992. A Guide to Task Analysis: The Task Analysis Working Group. CRC press.

KNKT, 2019. PT. Lion Mentari Airlines Boeing 737-8 (MAX); PK-LQP Tanjung Karawang, West Java Republic of Indonesia 29 October 2018. Accident Report KNKT.18.10.35.04. KOMITE NASIONAL KESELAMATAN TRANSPORTASI.

Kobayashi, H., Onoue, H., n.d. Brittle Fracture of Liberty Ships. `http://www.sozogaku.com/fkd/en/hfen/HB1011020.pdf`. Retrieved on Nov. 15th 2019.

Kobayashi, H., Terada, H., n.d. Mid-air Explosion of Comet I over the Mediterranean Sea. `http://www.shippai.org/fkd/en/hfen/HB1071012.pdf`. Retrieved on Nov. 15th 2019.

Konstandinidou, M., Nivolianitou, Z., Kiranoudis, C., Markatos, N., 2006. A fuzzy modeling application of CREAM methodology for human reliability analysis. Reliability Engineering & System Safety 91, 706–716.

Kuipers, B., 1986. Qualitative simulation. Artificial intelligence 29, 289–338.

Kuipers, B., 1994. Qualitative Reasoning: Modeling and Simulation With Incomplete Knowledge. MIT press.

Kuipers, B., 2001. Qualitative Simulation. Encyclopedia of physical science and technology 3, 287–300.

Lee, J., Chung, H., 2018. A new methodology for accident analysis with human and system interaction based on FRAM: Case studies in maritime domain. Safety Science 109, 57–66.

Lee, W.S., Grosh, D.L., Tillman, F.A., Lie, C.H., 1985. Fault Tree Analysis, Methods and Application — A Review. IEEE transactions on reliability 34, 194–203.

Leveson, N., 2004. A new accident model for engineering safer systems. Safety science 42, 237–270.

Lind, M., 2011. An introduction to multilevel flow modeling. Nuclear safety and simulation 2, 22–32.

Liu, Y.X., Lin, Y., Zhou, Y., 2017. 2D cellular automaton simulation of hot deformation behavior in a Ni-based superalloy under varying thermal-mechanical conditions. Materials Science and Engineering: A 691, 88–99.

Lu, L., Chan, C.Y., Wang, J., Wang, W., 2017. A study of pedestrian group behaviors in crowd evacuation based on an extended floor field cellular automaton model. Transportation Research Part C: Emerging Technologies 81, 317–329.

Mitchell, M., 2009. Complexity: A guided tour. Oxford University Press.

Nakao, M., n.d. Collapse of Tacoma Narrows Bridge. `http://www.sozogaku.com/fkd/en/hfen/HA1000632.pdf`. Retrieved on Nov. 15th 2019.

Neisser, U., 1976. Cognition and Reality. W.H.Freeman.

Neumann, J., Burks, A.W., et al., 1966. Theory of Self-reproducing Automata. volume 1102024. University of Illinois Press Urbana.

NTSB, 2010. Loss of Thrust in Both Engines After Encountering a Flock of Birds and Subsequent Ditching on the Hudson River US Airways Flight 1549 Airbus A320-214, N106US Weehawken, New Jersey January 15, 2009. Accident Report AAR-10/03. National Transportation Safety Board.

NTSB, 2014. Descent Below Visual Glidepath and Impact With Seawall Asiana Airlines Flight 214 Boeing 777-200ER, HL7742 San Francisco, California July 6, 2013. Accident Report AAR1401. National Transportation Safety Board.

Patriarca, R., Di Gravio, G., Costantino, F., 2017. A Monte Carlo evolution of the Functional Resonance Analysis Method (FRAM) to assess performance variability in complex systems. Safety Science 91, 49–60.

Petri, C.A., 1966. COMMUNICATION WITH AUTOMATA. Technical Report RADC-TR-65-377. Griffiss air force base, New York.

Qian, Y., Zeng, J., Wang, N., Zhang, J., Wang, B., 2017. A traffic flow model considering influence of car-following and its echo characteristics. Nonlinear Dynamics 89, 1099–1109.

Raabe, P., Houghton, W., Joksimovic, V., 1976. HTGR accident initiation and progression analysis status report. Volume 1. Introduction and summary. Technical Report. General Atomic Co., San Diego, Calif.(USA).

Rasmussen, J., 1986. Information Processing and Human-Machine Interaction. An approach to cognitive engineering .

Rasmussen, J., Pejtersen, A.M., Goodstein, L.P., 1994. Cognitive Systems Engineering. Wiley.

Rasmussen, J., Svedung, I., 2000. Proactive Risk Management in a Dynamic Society. Swedish Rescue Services Agency.

Reason, J., 2000. Human error: models and management. Bmj 320, 768–770.

Reason, J.T., Carthey, J., de Leval, M.R., 2001. Diagnosing "vulnerable system syndrome": an essential prerequisite to effective risk management. BMJ Quality & Safety 10, ii21–ii25.

Ross, D.T., 1977. Structured analysis (sa): A language for communicating ideas. IEEE Transactions on software engineering , 16–34.

Roth, E.M., Mumaw, R.J., 1995. USING COGNITIVE TASK ANALYSIS TO DEFINE HUMAN INTERFACE REQUIREMENTS FOR FIRST-OF-A-KIND SYSTEMS, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE Publications Sage CA: Los Angeles, CA. pp. 520–524.

Ruan, X., Zhou, J., Tu, H., Jin, Z., Shi, X., 2017. An improved cellular automaton with axis information for microscopic traffic simulation. Transportation Research Part C: Emerging Technologies 78, 63–77.

Saaty, T.L., 1990. How to make a decision: the analytic hierarchy process. European journal of operational research 48, 9–26.

SAE, 2016. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. SAE International .

Sarter, N.B., Woods, D.D., 1995. How in the World Did We Ever Get into That Mode? Mode Error and Awareness in Supervisory Control. Human Factors The Journal of the Human Factors and Ergonomics Society 37, 5–19.

Sarter, N.B., Woods, D.D., Billings, C.E., et al., 1997. Automation Surprises. Handbook of Human Factors and Ergonomics 2, 1926–1943.

Sawaragi, T., Matsubara, A., Kawakami, H., Horiguchi, Y., 2018. Artifact Design. Kyoritsu Shuppan.

Sheridan, T.B., 1992. Telerobotics, Automation, and Human Supervisory Control. MIT press.

Shioya, M., Mori, J., Ito, K., Mizutani, Y., Torikai, K., 2015. Atsuita Seizou Kouki no Kakuritsu Moderu to Seizou Hyoujun Kouki Sanshutsu Gijutsu no Kaihatsu (Development of Stochastic Model of Production Time and Estimation Technology for Standard Production Time in a Plate Mill). Tetsu to Hagane (Iron and Steel) 101, 574–583. doi:10.2355/tetsutohagane.TETSU-2015-024.

Simmon, D.A., 1998. Boeing 757 CFIT Accident at Cali, Colombia Becomes Focus of Lessons Learned. Flight Safety Digest 17(5/6), 1–31.

Slater, D., 2017. MODELLING, MONITORING, MANIPULATING AND MANAGING? Modelling Process Flow in Complex Systems.

Smith, D., Johnson, N.F., 2004. Evolution Management in a Complex Adaptive System: Engineering the Future. arXiv preprint cond-mat/0409036 .

Smith, K., Hancock, P.A., 1995. Situation Awareness Is Adaptive, Externally Directed Consciousness. Human factors 37, 137–148.

Stamatelatos, M., Vesely, W., Dugan, J., Fragola, J., Minarick, J., Railsback, J., 2002. Fault Tree Handbook with Aerospace Applications. Handbook. NASA.

STEEL, N., 2019. Steel Plates. `http://www.nssmc.com/product/catalog_download/pdf/A001en.pdf`. Retrieved on Oct. 18th 2019.

Swain, A.D., Guttmann, H.E., 1983. Handbook of human-reliability analysis with emphasis on nuclear power plant applications. Final report. Technical Report. Sandia National Labs.

Talmy, L., 1988. Force dynamics in language and cognition. Cognitive science 12, 49–100.

Ung, S.T., 2015. A weighted CREAM model for maritime human reliability analysis. Safety Science 72, 144–152.

Vanderhaegen, F., 2016. A rule-based support system for dissonance discovery and control applied to car driving. Expert Systems With Applications 65, 361–371.

Vanderhaegen, F., 2017. Towards increased systems resilience: new challenges based on dissonance control for human reliability in Cyber-Physical&Human Systems. Annual Reviews in Control 44, 316–322.

Vesely, W.E., Goldberg, F.F., Roberts, N.H., Haasl, D.F., 1981. Fault Tree Handbook. Handbook NUREG-0492. U.S. Nuclear Regulatory Commission.

Vicente, K.J., 1999. COGNITIVE WORK ANALYSIS: Toward Safe, Productive, and Healthy Computer-Based Work. CRC Press.

Vygotsky, L.S., 1980. MIND IN SOCIETY: The Development of Higher Psychological Processes. Harvard university press.

Weinberg, G.M., Weinberg, D., 1979. On the Design of Stable Systems. New York: Wiley.

Woods, D.D., Christoffersen, K., 2002. Balancing practice-centered research and design. Cognitive Systems Engineering in Military Aviation Domains. Wright-Patterson AFB, OH: Human Systems Information Analysis Center , 121–136.

Woods, D.D., Dekker, S., 2000. Anticipating the effects of technological change: a new era of dynamics for human factors. Theoretical issues in ergonomics science 1, 272–282.

Woods, D.D., Roth, E., Pople, H.E., 1988. Modeling Human Intention Formation for Human Reliability Assessment. Reliability Engineering & System Safety 22, 169–200.

Yang, Q., Tian, J., Zhao, T., 2017. Safety is an emergent property: Illustrating functional resonance in Air Traffic Management with formal verification. Safety Science 93, 162–177.

Yang, Z., Bonsall, S., Wall, A., Wang, J., Usman, M., 2013. A modified CREAM to human reliability quantification in marine engineering. Ocean engineering 58, 293–303.

Yoshida, Y., Ootani, M., Fjita, Y., 2002. Ninchiteki Bunnsekimoderu wo Mochiita Accident Management no Ishikettei Kagoritsu no Teiryouka (Quantification of a Decision-making Failure Probability of the Accident Management Using Cognitive Analysis Model). INSS journal 9, 183–204.

Yoshizawa, A., Oba, K., Kitamura, M., 2016. Lessons Learned from Good Practices During the Accident at Fukushima Daiichi Nuclear Power Station in Light of Resilience Engineering. IFAC-PapersOnLine 49, 245–250.

Zhou, Q., Wong, Y.D., Xu, H., Van Thai, V., Loh, H.S., Yuen, K.F., 2017a. An enhanced CREAM with stakeholder-graded protocols for tanker shipping safety application. Safety Science 95, 140–147.

Zhou, T., Wu, C., Zhang, J., Zhang, D., 2017b. Incorporating CREAM and MCS into fault tree analysis of LNG carrier spill accidents. Safety Science 96, 183–191.

研究業績

学術雑誌(紀要・論文集等も含む）に発表した論文

[1] Takayuki Hirose, Tetsuo Sawaragi, Yukio Horiguchi: Safety Analysis of Aviation Flight-Deck Procedures Using Systemic Accident Model; IFAC Proceedings Volumes, Elsevier Ltd., Volume 49, Issue 19, pp. 19–24, 2016.
（下記国際会議での発表 [1]の内容が，審査を経て Proceedings Volume に掲載され発刊されたもの）

[2] Takayuki Hirose, Tetsuo Sawaragi, Yukio Horiguchi, Hiroaki Nakanishi: Safety Analysis for Resilient Complex Socio-Technical Systems with an Extended Functional Resonance Analysis Method; Int J Astronaut Aeronautical Eng 2:012, 2017（査読あり；採択済）

[3] Takayuki Hirose, Tetsuo Sawaragi: Development of FRAM Model Based on Structure of Complex Adaptive Systems to Visualize Safety of Socio-Technical Systems; IFAC Proceedings Volumes, Elsevier Ltd., Volume 52, Issue 19, pp. 13–18, 2019.
（下記国際会議での発表 [6]の内容が，審査を経て Proceedings Volume に掲載され発刊されたもの）

[4] Takayuki Hirose, Tetsuo Sawaragi: Extended FRAM model based on cellular automaton to clarify complexity of socio-technical systems and improve their safety; Safety Science, Elsevier Ltd., Volume 123, 2020. （査読あり；採択済）

[5] Takayuki Hirose, Tetsuo Sawaragi, Hideki Nomoto, Yasutaka Michiura: Functional Safety Analysis of SAE Conditional Driving Automation in Time-Critical Situations and Proposals for Its Feasibility; Cognition, Technology & Work, Springer. （査読あり；投稿中）

国際会議における発表（発表者は下線で表示）

[1] Takayuki Hirose, Tetsuo Sawaragi, Yukio Horiguchi: Safety Analysis of Aviation Flight-Deck Procedures Using Systemic Accident Model; The 13th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems, Kyoto Japan, August 2016. （口頭・査読あり）

[2] Takayuki Hirose: Safety Analysis of Standard Operation Procedures Under the Influence of Variabilities （邦題：ゆらぎの影響下で遂行される標準作業手順の脆弱性解析）; Students Research Exchange Workshop in Design of International Design Symposium in Kyoto 2018, Kyoto Japan, March 2018.（ポスター・査読なし）

[3] Takayuki Hirose, Tetsuo Sawaragi, Yukio Horiguchi: Numerical Simulation of Complex Supply-Chain Systems with an Extended Model of Functional Resonance Analysis Method; Third International Workshop on Functional Modelling for Design and Operation of Engineering Systems, Kurashiki Japan, May 2018. （口頭・査読あり）

[4] Takayuki Hirose, Tetsuo Sawaragi, Yukio Horiguchi: Numerical Safety Analysis of Complex Supply-Chain Systems Integrating Functional Resonance Analysis Method and Cellular Automaton; The 12th FRAMily meeting/workshop, Cardiff Wales UK, June 2018. （口頭・査読あり）

[5] Takayuki Hirose, Tetsuo Sawaragi: FRAM and Complex Adaptive Systems: New Model to Simulate Complex Dynamics of Socio-Technical Systems; The 13th FRAMily meeting/workshop, Malaga Spain, May 2019. （口頭・査読あり）

[6] Takayuki Hirose, Tetsuo Sawaragi: Development of FRAM Model Based on Structure of Complex Adaptive Systems to Visualize Safety of Socio-Technical Systems; The 14th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems, Tallinn Estonia, September 2019. （口頭・査読あり）

[7] Tetsuo Sawaragi, Yukio Horiguchi, Takayuki Hirose: Design of Productive Socio-Technical Systems by Human-System Co-Creation for Super Smart Society; The 21st IFAC World Congress, Berlin Germany, 2020,

国内学会・シンポジウムにおける発表（発表者は下線で表示）

[1] 広瀬貴之，椹木哲夫，堀口由貴男：ファジィ CREAM と FRAM の統合による操作手順脆弱性の系統的分析；日本知能情報ファジィ学会関西支部第 2 回学生研究発表会，大阪大学豊中キャンパス，2016 年 10 月（口頭・査読なし）

[2] 広瀬貴之，椹木哲夫，堀口由貴男，中西弘明：機能共鳴解析のためのファジィ推論に基づく作業手順の脆弱性評価；第 61 回システム制御情報学会研究発表講演会，京都テルサ，2017 年 5 月（口頭・査読なし）

[3] 広瀬貴之，椹木哲夫，堀口由貴男，中西弘明：機能共鳴解析手法の拡張による社会・技術システムのレジリエンス強化のための安全解析；第 145 回ヒューマンインタフェース学会研究会，　東北大学青葉山東キャンパス，2017 年 6 月（口頭・査読なし）

[4] 広瀬貴之，岩見一樹，椹木哲夫，堀口由貴男：機能共鳴解析手法を用いた生産工程サプライチェーンのゆらぎ解析；第 62 回システム制御情報学会研究発表講演会，京都テルサ，2018 年 5 月（口頭・査読なし）

[5] 広瀬貴之，椹木哲夫：レジリエントな社会・技術システムのデザインのための拡張型機能共鳴解析手法を用いた鉄鋼生産ラインのゆらぎ解析；日本鉄鋼協会第 176 回秋季講演大会，東北大学川内キャンパス，2018 年 9 月（口頭・査読なし）

[6] 広瀬貴之，椹木哲夫：A Quantitative Approach of FRAM: How to Implement the Numerical Simulation；FRAM 勉強会 2018，IPA: 独立行政法人 情報処理推進機構，2018 年 11 月（口頭・査読なし）

[7] 広瀬貴之，椹木哲夫：機能共鳴分析手法（FRAM）に基づく社会・技術システムの安全評価シミュレータの開発；平成 30 年シンビオ社会研究会 第 2 回研究談話会，KRP，2018 年 12 月（口頭・査読なし）

[8] 広瀬貴之，椹木哲夫：Using Unstable for Stable: Examples to Validate Why Things Go Right；AI/IoT のための安全性シンポジウム，国立情報学研究所，2019 年 11 月（口頭・査読なし）

受賞歴等

[1] 平成 29 年度 京都大学大学院工学研究科 馬詰研究奨励賞受賞（副賞：海外研究渡航費用補助），2017 年 7 月

[2] 平成 30 年度京都大学デザイン学大学院連携プログラム研究助成費 公募採択，2018 年 4 月