MIT ENGINEERING SYSTEMS LAB

Safety III: A Systems Approach to Safety and Resilience

Prof. Nancy Leveson

Aeronautics and Astronautics Dept., MIT

7/1/2020

Abstract: Recently, there has been a lot of interest in some ideas proposed by Prof. Erik Hollnagel and labeled as "Safety-II" and argued to be the basis for achieving system resilience. He contrasts Safety-II to what he describes as Safety-I, which he claims to be what engineers do now to prevent accidents. What he describes as Safety-I, however, has very little or no resemblance to what is done today or to what has been done in safety engineering for at least 70 years. This paper describes the history of safety engineering, provides a description of safety engineering as actually practiced in different industries, shows the flaws and inaccuracies in Prof. Hollnagel's arguments and the flaws in the Safety-II concept, and suggests that a systems approach (Safety-III) is a way forward for the future.

Safety III: A Systems Approach to Safety and Resilience

Contents

Figures

Preface

Recently, there has been a lot of interest in some ideas proposed by Prof. Erik Hollnagel and labeled as "Safety-II" and argued to be the basis for achieving system resilience. He contrasts Safety-II to what he describes as Safety-I, which he claims to be what engineers do now to prevent accidents. What he describes, however, has very little or no resemblance to what is done today or to what has been done in safety engineering for at least 70 years.

First, should you take my word for this? I have worked in safety engineering for 40 years. Here's a little of my relevant background. I have degrees in mathematics, management, and computer science and did graduate work in cognitive and behavioral psychology. I have written two books on system safety (*Safeware* [Leveson, 1995] and *Engineering a Safer World* [Leveson, 2012]) and hundreds of papers on the topic. My efforts have been rewarded with many awards, most recently an IEEE Medal for Environmental and Safety Technologies. I am an elected member of the National Academy of Engineering.

I also am fascinated by engineering history and have read much about how engineers handled safety for the past hundred or so years. In practice, I have worked in almost all aspects of aerospace and defense and, to a lesser extent, nuclear power, petrochemicals, patient safety and medical devices, most forms of transportation (particularly aircraft and automobiles), etc. I have also participated in writing some major accident reports (Deep Water Horizon, the Columbia Space Shuttle, and Texas City) and many less well-known ones. Finally, in the past few years, I have been encouraged to look into workplace safety because people felt that the engineering approaches that I have created might be useful there.

I provide this background because I don't recognize Prof. Hollnagel's definition of Safety-I in my 40 years of experience in safety engineering. It is just *not* what is done in practice except, perhaps, in a very few organizations with the least sophisticated safety approaches. His analysis also confuses the almost totally different fields of product/system safety and workplace safety.

Prof. Hollnagel tears apart his strawman Safety-I and recommends an alternative, which he calls Safety-II. In my experience, again, Safety-II is a giant step backward, particularly if it takes resources and attention away from more successful approaches. It contains the types of practices used in the past, mostly very long ago but also more recently in industries that have many accidents and that usually blame them all on the human operators. These practices have led to many unnecessary deaths and injuries.

The Safety-II approach was rejected long ago in sophisticated engineering projects because it is not effective. Goals such as resilience, flexibility, and adaptability are important, but they are much more likely to be achieved using approaches other than Safety-II. These properties must be built into the system as a whole—they are not a function simply of the behavior of human operators, which seems to be the almost total emphasis in Safety-II. There certainly are a few aspects of Safety-II that might be useful in limited ways, but following the overall approach, I believe, is likely to lead to unnecessary accidents.

In this paper, I explain these very strong statements and note that Prof. Hollnagel and his followers seem unaware of the successful use of a systems approach to safety, which is called "System Safety"[1] by its practitioners. They may not know about it; it was developed and used primarily in the United States.

System Safety was created for and has been used over the past 70 years in aerospace and defense to cope with the most dangerous systems being created. In my work and writings, I have extended this very successful practice to handle the evolution (and sometimes revolutionary change) of engineering

[1] The term "system safety" has been adopted recently as a general term for safety engineering by people not familiar with the special field of System Safety developed long ago. I will differentiate them here by using capital letters to denote the specialized field of System Safety.

practices over time. These changes include greatly increasing complexity, the extensive and growing use of computers and other forms of new technology, and a changing role of humans in complex systems. In this paper, I call this general approach *Safety-III* to put it into the Hollnagel context. It is not new, however—the general practices have been around for a very long time, but primarily used in the most sophisticated and sometimes secretive engineering contexts. It can provide a template for advances in *all* industries, including product/system safety and workplace safety, going forward. Changes and advances will be needed to keep it relevant for engineering in the future, of course, as our technology and society change.

One of the dangers of critiquing someone's approach is determining exactly what that approach is. Our views evolve over time as more is learned, and we all change them in small or even major ways with more experience. In addition, many people write papers about someone else's concept and interpret it differently than the original author, introducing their own slant and representing their own experiences. I've seen this in papers by the proponents of Safety-II other than Prof. Hollnagel, particularly in healthcare safety. To try to stay as close as possible to the original conceptions of Safety-I and Safety-II, as defined by Prof. Hollnagel, in this paper I use only the writings of Prof. Hollnagel himself, basically his two books *Safety I and Safety-II: The Past and Future of Safety Management* [Hollnagel, 2014] and *Safety-II in Practice* [Hollnagel, 2018]. Any misunderstandings about what was written in these sources stem from my own shortcomings, although I have invested a large amount of effort to try to understand these books.

Some misunderstandings may arise because my experience in safety engineering over the past 40 years is so different than what is described in the Hollnagel writings that I find it difficult to relate what I read to what I have seen in actual practice. In addition, Prof. Hollnagel does not use engineering terminology and almost never carefully defines the terms he does use nor use them in a consistent way throughout. The combination of vague terminology, along with almost no examples or citations (especially to engineering literature) for most of his assertions, made it very hard for me to relate his claims to my real-world experience. I will try to use lots of examples in this paper to assist the reader in understanding my explanations.

One of the most puzzling (and frustrating) parts of Prof. Hollnagel's first book is that he spends 200 pages tearing down some of the most successful practices in safety engineering today and then suggests at the end that it should all be done anyway, along with looking at what "goes right" (Safety-II). Creating and tearing apart this strawman was not necessary and creates misunderstandings about current practices as they actually exist that I feel I need to correct.

I could nitpick most of what Prof. Hollnagel writes—there are technical mistakes on nearly every page—but I will concentrate on what seem to me to be the major misunderstandings, along with providing careful definitions of standard terminology and concepts in this field to assist all of us in better communication.

Along with critiquing Safety-I and Safety-II, what I am here calling Safety-III is briefly described. In the broader safety engineering community, Safety-III is described as the "systems approach" to safety, i.e., System Safety. As I said, it is not new: it has been around since the 1950s. I also briefly describe my own extensions to this approach in the past 25 years and the directions I think we need to go in the future.

Does "Safety-I" Exist?

It is not reasonable to characterize whole groups and industries as having one approach to safety, as is done in Prof. Hollnagel's misleading description of Safety-I—each industry and group within it are very different. There is no one monolithic approach in our attempts to reduce losses: The approach to safety varies with each industry and sometimes individual companies in an industry. However, I have not come across *any* industries that approach safety in the way that Prof. Hollnagel describes as Safety-I.

An aside: **Strawman arguments**. Because Prof. Hollnagel uses the strawman fallacy so much in these books, indeed it is the basis for the entire argument in his 2014 book, it is useful to review this type of logical fallacy. A straw man is a form of argument and an informal fallacy based on giving the impression of refuting an opponent's argument, while actually refuting an argument that was not presented by that opponent. One who engages in this fallacy is said to be "attacking a straw man". The basic form of such an argument is to start with a misrepresentation of someone's argument to make it easier to attack and make your own position seem more reasonable. Here is an example from Hollnagel [2014, page 44):

> *The usual practice of both accident analysis and risk assessment implies a view of safety according to which systems work because the following conditions are met:*
>
> - *Systems are well designed, meaning that the functions of all the components integrate smoothly to produce the desired outcome. It is in particular assumed that system designers have been able to foresee everything that can go wrong and to take the necessary precautions (such as redundancy, graceful degradation, barriers, and defences).*
> - *…*

This statement argues that practicing accident analysis and risk assessment means that someone (most people?, everyone?) believes that systems "*work*" (which is undefined and meaningless) because they are well designed and that the designers have been able to "foresee everything that can go wrong" and take the necessary precautions.

First, in 40 years, I have never met an engineer or safety engineer that believed that everything can be foreseen and prevented and this is particularly true of those engaged in accident analysis. In fact, there is a common phrase used in engineering where engineers talk about the "unknown unknowns." It is a long and illogical stretch from the fact that engineers try to prevent accidents to implying they believe they can do a perfect job. The same argument can be applied for risk assessment. If risk assessment assumed that everything was foreseen and the necessary precautions had been taken, then the risk should be zero and there is no need for an assessment.

This type of argument runs through the book yet there are no examples, no citations, and only vague suggestions about who might be the people who believe these things. It is stated as obviously true without any attempt at proof. The unproven assertion is then used to "prove" that Safety-II is better.

Here is another version found in a different place in the book where Prof. Hollnagel describes what he calls "common myths":

> Page 87[2]*: Belief that all accidents are preventable---all accidents have causes, and causes can be found and eliminated*

There is no hint about who believes this (the average citizen, psychologists, engineers, social scientists?) nor any evidence provided that *anyone* believes it, let alone labeling it as a "common belief." While it is probably true that most people believe that all accidents have causes, it is not necessarily true that everyone believes that the causes can always be found (e.g., Malaysian Airlines 370), but for the most part this statement is true. More important, however, even if they can be found, it does not mean they can be eliminated. I can't think of anyone in engineering who believes the last phrase, i.e., that all accident causes can be eliminated. It implies that engineers are incredibly naïve. Where are some citations? Examples? Stating something is a "common myth" without any proof that anyone believes it is a type of strawman argument.

Throughout the books, Prof. Hollnagel ascribes ridiculous beliefs to engineers. For example, on Page 42, he says that machines can be formally defined using a finite-state automaton (inputs, outputs, internal states, and state transitions). Actually, this is only true for computers or finite state machines,

[2] Throughout this paper, a page number alone will refer to Prof. Hollnagel's first book on Safety-II published in 2014. If the reference is to his second book, (2) will appear after the page number, e.g., page 47(2).

not all machines. He then concludes "*But such a description also forces designers (and managers) to think of humans in the same way.*" Why would this be true? I know the difference between humans and computers, and I would guess that is true for virtually all engineers and managers.

> Page 42*: "The purpose of the [system] design, combined with suitable training when necessary, is to make the user respond as a finite automaton, and it is not far from that to actually think about the user as an automaton."*

Where are some citations, examples, etc. of engineers thinking this way? Why would the purpose of system design be to make the user respond as a finite automaton? If that were the real goal, then the users would be replaced by computers. The reality is exactly the opposite, i.e., engineers usually think the humans will *not* act like a machine. For example, many designers and engineering standards assume that the human operators will be able to detect and intervene to fix the errors and failures in the physical and software parts of the process being controlled. But note that Prof. Hollnagel does not actually say that engineers think this way. He simply says "*it is not far from that to actually think*" that way. He implies that engineers think this way without actually having to say so. But he does say exactly that later in the book:

> Page 81: *To an engineer, humans are system components whose successes and failures can be described in the same way as for equipment*.

Where are some citations or examples of engineers doing this? Once again, it is a useful strawman to entice the reader to accept Prof. Hollnagel's conclusions about Safety-II and vilify what he claims that engineers do today. The book has almost no examples of accidents, no references (particularly engineering references), or any other support for these claims.

Strawman arguments, which comprise most of the arguments in this book, are ones that exaggerate, misrepresent, or just completely fabricate what are supposedly someone else's beliefs. It is then much easier to present your own position as being reasonable and, in fact, the only reasonable alternative.

Here is an example of a different type of strawman argument or logical fallacy used in Prof. Hollnagel's books where something is claimed to imply something that it doesn't.

> "After Will said that we should put more money into health and education, Warren responded that he was surprised that Will hates our country so much that he wants to leave it defenseless by cutting military spending".[3]

An equivalent example is found on page 58 of Prof. Hollnagel's first book:

> Page 58*: Because Safety-1 focuses on preventing adverse outcomes, safety competes with productivity*.

The first part of this statement *"Because Safety-I focuses on preventing adverse outcomes"* (e.g. Will wants to spend more money on health and education) does not imply the second part "*Safety competes with productivity*" (Will wants to cut military spending). Where is an argument or evidence that a focus on preventing accidents makes systems less productive?

In fact, preventing accidents usually leads to increased productivity as accidents can have tremendously negative impacts on productivity. The most effective measures used to prevent accidents do not decrease productivity, they increase it. When operators do not have to worry about their actions causing a tragedy, they can usually place more emphasis on efficiency. But Prof. Hollnagel, by setting up the fallacious argument, comes up with an argument for his desired conclusion, i.e., because Safety-II does not focus on preventing adverse outcomes, it does not compete with productivity—which is the exact same form of fallacious argument. There certainly are some who believe that safety competes

[3] https://yourlogicalfallacyis.com/strawman

with productivity, but Prof. Hollnagel states this belief as a fact that decreased productivity results from preventing adverse outcomes, without explaining why.

Throughout the book, Prof. Hollnagel makes what sound like logical arguments but are not; they represent fallacious reasoning or sophistry.

One of the most important straw men in the book is the description of Safety-I. This simply is not how safety is treated by most all engineering organizations. Unlike Prof. Hollnagel's simplification of all of safety engineering as "Safety-I", there is in fact no single approach to engineered system safety that exists: a great variety of solutions are used, depending on a number of factors including differences in the problems that need to be solved, different cultures among industries, and different histories.

Instead of starting from Prof. Hollnagel's greatly oversimplified description of Safety-I that makes what is currently done look silly, we need to start with examining the whole spectrum of what people actually do or what has been suggested they do to prevent accidents. This information will make it possible to determine what works for what types of systems, what does not work and why, and how we should go forward to improve all aspects of safety.

There are basic differences in different industries, different communities (for example, workplace/occupational safety vs. safety of engineered systems), different parts of the world (cultures and political structures), and different disciplines (social scientists vs. engineers).[4] Much of Prof. Hollnagel's first book describes workplace safety and an assumption is made that workplace safety and product/system safety are somehow related or approached in similar ways. Nothing could be farther from the truth.

Only a brief overview is included here as a complete description would fill a long book. This section surveys some of these different approaches to safety engineering and their historical development. After the survey, Prof. Hollnagel's specific claims about Safety-I are examined in more detail.

The biggest difference is between workplace safety and product/system safety. So let's start there and then look at the differences between how product/system safety is tackled within different types of industries.

Differences between Workplace Safety and Product/System Safety

Prof. Hollnagel is a social scientist, with degrees in psychology, and seems to have had limited experience with safety engineering, judging from the contents of the two books being used here. He does not distinguish between the approaches to workplace and product/system safety, and general statements supposedly about both actually focus solely on workplace safety history, concepts and approaches. Most of his statements about what safety engineering is, its history, and what safety engineers believe are completely wrong.

The first important problem with Prof. Hollnagel's argument is that he lumps workplace safety and product/system safety together. He describes the limitations of the practices used in workplace safety and assumes that the same limitations exist in product/system safety. In fact, the practices in product/system safety are completely different and also differ among industries. They are not related except in some historical ways, which are briefly described below.

Note that almost all the references (of the very few included) in these two Hollnagel books are to social scientists and their writings. The vast literature on safety engineering or engineering in general,

[4] One omission is important to acknowledge here. Services can also be safety critical, but the emphasis in this paper will be on engineered systems and products as well as workplace safety. Complex systems that provide services (which is most of them), such as air traffic control or healthcare, are included but individual service providers, such as cosmeticians or electricians are not. Individual service providers are usually covered by standard codes of practice and are often licensed by the government.

with only a very few exceptions, is mostly ignored. My view of workplace safety may, of course, similarly be distorted in the following as my experience in that field is more limited than in product/system safety. In fact, Prof. Hollnagel does a good job of describing current practices and some aspects of the history of workplace safety. I will try to fill in this history, particularly in the instances in which historically workplace and product/system safety have overlapped.

Workplace safety is sometimes also called Occupational Health and Safety, Industrial Safety, or Industrial Hygiene. It focuses on injuries to workers during, and as a result of, their job responsibilities, such as exposure to industrial chemicals and toxins, human injury due to strain or stress involved in their job (ergonomics), or other workplace injuries such as falling off ladders. A great emphasis in workplace safety has been placed on psychology—particularly behavioral psychology—in preventing workplace injuries using training, reward and punishment, and various other types of persuasion to encourage workers to follow rules and procedures and to wear personal protective equipment (PPE). The foundational assumption appears to be that worker injuries are due to the workers themselves and if they just were more careful and followed instructions, injuries would not occur.

Much less emphasis has been on changing the design of the system in which the humans work, although guards or barriers and other types of simple devices may be used in the design of assembly and production line automation and have been a part of workplace safety since at least the 19th century.

When an accident does occur, workers are usually blamed, with the most common cause cited being non-compliance with specified procedures. On this, I wholeheartedly agree with Prof. Hollnagel. The problem is that he seems to believe that this same approach is used in product/system safety. The role of compliance with procedures in complex engineered systems is discussed in detail later in this paper.

In contrast to workplace safety, product safety or system safety or safety engineering (all of which are terms used in various contexts) focuses on the dangers of the products/systems we design and use. There is clearly some overlap with workplace safety—for example, workers may be injured because of the dangers of the engineered products they use, e.g., by machines or robots in factories. In general, however, workplace safety concentrates on how employees or workers perform their work while product safety focuses on the characteristics of engineered products. Historically there has been little overlap of practices or practitioners in these two areas of safety for the past 100 years, even when the problem is that of workers being injured by the use of engineered products in the workplace. Most of the descriptions of Safety-I in Prof. Hollnagel's book appear to come from workplace safety.

Prof. Hollnagel in his first book spends a great deal of time talking about such workplace safety topics as Heinrich, Heinrich's Triangle, Heinrich's model of accident causation (called the Domino model), Taylorism, and behaviorism. None of these have anything to do with product/system safety. There is a puzzling number of pages devoted to discussing Taylorism, which contains nothing particularly relevant to either type of safety.

To understand the differences, the next section first describes the history of both and their overlaps and divergences. Then I provide what is an engineer's view of workplace safety and product/system safety today.

Workplace and Product/System Safety History

There are two relevant historic narratives about safety, one focusing on the legal aspects and one on the technical ones.[5]

[5] Much of the following section contains excerpts from my book, *Safeware*, published by Addison-Wesley in 1995.

A Brief Legal View of the History of Safety

Humans have always been concerned about their safety. Prior to the industrial age, natural disasters provided the biggest challenge. Things started to change in the early part of the Industrial Revolution in Europe and the United States. Workers in factories were considered expendable and were often treated worse than slaves: slaves cost a great deal of money and owners wanted to protect their investment, but workers cost nothing to hire or replace [Gloss and Wardle, 1984]. The prevailing attitude was that when people accepted employment, they also accepted the risks involved in the job and should be smart enough to avoid danger. At the same time, factories were filled with potentially dangerous equipment, such as unguarded machines, flying shuttles, and open belt drives, as well as unsafe conditions, such as open holes. There were no fire escapes, and the lighting was inadequate. Hardly a day went by without some worker being maimed or killed [Gloss and Wardle, 1984].

Without workers' compensation laws, employees had to sue and collect damages for injuries under common law. In the United States, the employer almost could not lose because common law precedents established that an employer did not have to pay injured employees if:

1. The employee contributed at all to the cause of the accident: C*ontributory negligence* held that if the employee was responsible or even partly responsible for an accident, the employer was not liable,
2. Another employee contributed to the accident: The *fellow-servant doctrine* held that an employer was not responsible if an employee's injury resulted from the negligence of a co-worker, and
3. The employee knew of the hazards involved in the accident before the injury and still agreed to work in the conditions for pay: The *assumption-of-risk* doctrine held that an injured employee presumably knew of and accepted the risks associated with the job before accepting the position.

It is not, therefore, surprising that most investigations of accidents, which were conducted by management, found that the workers were responsible.

A non-employee fared somewhat better because a stranger presumably had no knowledge of the potential hazards in the plant. However, under the doctrine of *respondeat superior*, the employer was not responsible for an injury to a third party caused by an employee, and the employee usually did not have the means to compensate the injured party [Gloss and Wardle, 1984; Petersen, 1971].

The horrible conditions existing at this time led to social revolt by activists and union leaders. Miners, railroad workers, and others became concerned about the hazards of their jobs and began to agitate for better conditions. Voluntary safety organizations, such as the American Public Health Association, the National Fire Protection Association, and Underwriters Laboratories, were formed in the late nineteenth century and were active in setting standards. The first efforts focused on health rather than safety: accidents were seen as fortuitous events over which we had little control [Ferry, 1984].

Concern in Europe over worker safety preceded that in the United States. Otto von Bismarck established workers' compensation and security insurance, paid for by the employees, in Germany during the 1880s. Bismarck sought to undercut the socialists by demonstrating to the German working class that its government was in favor of social reform [Gloss and Wardle, 1984]. Soon, most countries in Europe followed Bismarck's lead.

Other types of safety legislation were also passed in Europe. For example, the Factory and Workshop Act of 1844 in Great Britain was the first legislation for protection from accidents involving shafts, belts, pulleys, and gears used to transmit power from a water wheel or a stationary steam engine throughout the factory. Later, laws setting standards for equipment such as high-pressure steam engines and high-voltage electrical devices were enacted.

The textile industry was one of the first to become mechanized and also contributed more than its share of injuries to workers. The abuses in the mills led to protective legislation and codes in many

countries by the end of the nineteenth century. The same was true for industries that employed metal and woodworking machines. Unfortunately, many safety devices, added only grudgingly, were poorly designed or ineffective [Roberts, 1984].

In the United States, employers remained indifferent to the many workers being killed and maimed on the job. Eventually, social revolt and agitation by unions against poor working conditions led to social reforms and government intervention to protect employees and the public.

Individual state laws preceded federal legislation. The first successful regulatory legislation in the U.S. was enacted in 1852, when Congress passed a law to regulate steam boat boilers. This law resulted from public pressure, enhanced by a series of marine disasters that had killed thousands of people.

In 1908, the state of New York passed the first workers' compensation law, which, in effect, required management to pay for injuries that occurred on the job regardless of fault. The New York law was held to be unconstitutional, but a similar law, passed in Wisconsin in 1911, withstood judicial scrutiny. All other states have now enacted similar laws, with the last being in 1947. Many other countries have similar laws.

When management found that it had to pay for injuries on the job, effort was put into preventing such injuries, and the organized industrial safety movement was born. Owners and managers also began to realize that accidents cost money in terms of lower productivity and started to take safety seriously. The first industrial safety department in a company was established in the early 1900s. The heavy industries, such as steel, in which workplace accidents occurred frequently, began to set safety standards.

This historical legal narrative was an important impetus to the history of technical developments. So now let's look at the historical narrative from a technical viewpoint.

A Technical View of the History of Safety

Safety has been an issue from early times and, indeed, the need for safe design is even mentioned in the Bible. In more modern times, however, a few engineers began to recognize the need to prevent hazards early in the industrial era, when the machines they were designing and building began to kill people. Watt warned about the dangers of high-pressure steam engines in the early 1800s. The Davey Safety Lamp, invented around the same time, helped decrease some of the danger of methane gas explosions in mines. In 1869, George Westinghouse developed a brake based on compressed air, which made railroad travel vastly safer for riders and crew. In 1868, the first patent was granted by the U.S. Patent Office for a machine safety device—an interlocking guard used in a machine for filling bottles with carbonated water. Other patents soon appeared for guards for printing presses, two-hand controls, and circular saws and other woodworking machines [Roberts, 1984].

Near the end of the nineteenth century, engineers began to consider safety, as well as functionality, in their designs instead of simply trying to add it on in the form of guards. While the use of guards was prevalent then and still is in workplace safety, much more sophisticated engineering solutions have been found and are used in complex systems, as will be seen later.

Many of the dangers involved in workplace accidents were due to the increasing use of engineered devices in the workplace so there was much overlap at this early time in the history of safety engineering and workplace safety. A divergence between the two fields occurred later, as you will see, including much more sophisticated design techniques in engineered systems beyond the use of simple guards and barriers.

One of the first organizations to study accidents was the Society for Prevention of Accidents in Factories (called the Mulhouse Society), which was founded in the town of Mulhouse in Alsace-Lorraine. The Mulhouse Society held annual meetings to exchange ideas about safety improvements in factories and published an encyclopedia of techniques for injury prevention in 1889 and 1895 [Mulhouse Society,

1895]. By the early part of the twentieth century, a German engineering society was established for the prevention of accidents [Roberts, 1984].

About the same time, the engineering technical literature started to acknowledge that safety should be built into a design. The first paper dealing with the safe design of machinery was presented to the American Society of Mechanical Engineers by John H. Cooper in 1891:

> *It is an easy task to formulate a plan of accident-preventing devices after the harm is done, but the wiser engineer foresees the possible weakness, as well as the dangers ahead, which are involved in his new enterprise, and at once embodies all the necessary means of safety in his original design* [Cooper, 1891, p.250]

In 1899, John Calder published a book in England, *The Prevention of Factory Accidents*, which provided accident statistics and described safety devices in detail. The book emphasized the need for anticipating accidents and building in safety and argued that legislation to compel manufacturers to provide safe products would probably not be necessary; free market forces would provide a stronger incentive.

> *Safeguarding by the user, of some kind, can in the long run be compelled by statute, but the author's experience is that, in the case of the multitude of occupiers of small factories, with no mechanical facilities or aptitude, nothing can take the place of good fencing fitted by the makers, and all accidents thereby <u>prevented</u> by being anticipated* (quoted in [Roberts, 1984, p. 89]).

Calder later moved to the United States and changed his mind: By 1911, he was calling for legislation to force manufacturers to provide safe products [Calder, 1911].

Trade journals started to editorialize about the need for designers to eliminate hazards from machinery:

> *We reiterate that the time to safeguard machinery is when it is on the drawing board; and designers should awaken fully to a sense of their responsibility in this respect. They should consider safety as of equal importance with operating efficiency, for if the machines unprotected, are not safe to work, they are failures, no matter how efficient they may be as producers* (quoted in [Roberts, 1984, p. 90]).

The first American technical journal devoted solely to accident prevention, *The Journal of Industrial Safety*, began publication in 1911, and individual states began to hold safety conferences. The first American National Safety Congress, organized by the Association of Iron and Steel Electrical Engineers, was held in Milwaukee in 1912.

In 1914, the first safety standards published in the United States, the Universal Safety Standards, were compiled under the direction of Carl M. Hansen [Hansen, 1914]. These standards required first defining hazards and then finding ways to eliminate them by design. This approach is the essence of safety engineering today; it is interesting how far back it goes.

Engineers at this time started to study safety as an independent topic. A study examined the widely held conviction that safer machines with guards were inefficient and resulted in reduced production. The study, conducted by a group that included the major engineering societies at the time, involved employees of 20 industries and 60 product groups who had a combined exposure of over 50 billion hours. The final report confirmed the hypothesis that production increased as safety increased—a lesson still be learned by many people today.

The study also explained the historical increase in accidents despite industrial safety efforts to reduce them. The increase was found to be related to the tremendous increase in the rate at which American industry was becoming mechanized. Mechanization affected safety in three ways: (1) it displaced the use of hand tools, (2) it increased the exposure of maintenance personnel to machine hazards, and (3) it allowed increased operating and material-feed speeds. The primary conclusion of the study was "*While […] there has been this recent increase in the hazard of industry per man-hour, production per man-hour*

has increased so much more rapidly that the hazard in terms of production has decreased" [Roberts, 1984].

Up to this point in time, there are two things to note:

1. Workplace safety and product safety were related and intertwined, primarily because many of the new engineered devices were used in workplaces and were created to reduce workplace injuries. At the same time, however, designing safety into more generally used hazardous products, such as the steam engine, that affected public safety and not just the workplace, was also a focus of study and engineering practices.
2. Unlike Prof. Hollnagel's description of Safety-I as being almost totally preoccupied with accident investigation, investigation was actually a very small part of safety engineering in the past (and now). Much more concern was placed on identifying hazards and designing safety into products from the beginning of development.

At this time (around the 1930s), workplace and product/system safety diverged and started to develop separately.

On the workplace safety side, in 1929, H.W. Heinrich, who worked for an insurance company and thus had access to a large number of accident reports, published a study of 50,000 industrial accidents, concluding that for every serious injury that occurred in the workplace, there were 20 minor injuries and 300 incidents involving reportable injury. He also suggested that thousands of near misses were occurring as well—although it is not clear how he got this information from the accident reports. This hypothesis became known as Heinrich's pyramid (or triangle) and is described by Prof. Hollnagel.

In 1931, Heinrich published a book , Industrial Accident Prevention, in which he claimed that workplace accidents result from unsafe actions and unsafe conditions but suggested that people cause far more accidents that do unsafe conditions [Heinrich, 1931].

Seizing on these arguments, opponents of mechanical (i.e., engineering) solutions to workplace safety began to direct attention away from unsafe machinery and toward unsafe user acts. Claims were made that accident-prone workers and carelessness were responsible for 85 to 90 percent of all industrial accidents. These arguments were based on the fact that accident investigations concluded that workers were responsible for most accidents and that accidents were occurring despite the use of machine guards. The arguments ignored the facts, of course, that (1) finding workers responsible for accidents was to the advantage of the companies conducting the investigations as that made them not responsible for the worker deaths and injuries and that (2) many accidents such as slips and falls or injuries from falling objects or from lifting were not machine-related and thus could not be prevented by any type of guarding [Roberts, 1984]. More recent data show that humans are more likely to be blamed for accidents than are unsafe conditions, even when unsafe conditions make human error almost inevitable [Leveson, 2012].

A more important question than assigning blame or cause may be how to eliminate accidents in the future. Hansen, back in 1915, had written

> *Forgetfulness, for example, is not a crime deserving of capital punishment; it is one of the universal frailties of humanity. The problem is, therefore, to destroy as far as possible the interrelationship between safety and the universal shortcomings, which can be done by designing the safeguards on machines and equipment so that if a man's acts are essential to safety, it becomes mechanically necessary for him to perform this act before proceeding with his task* [Hansen, 1915, p.14].

This design principle is one (of many) that is part of the basic approach to designing industrial machinery and processes today.

With this brief introduction to history, let's look at the state of workplace and then product/system safety today.

An Engineer's View of Workplace Safety

While I have degrees in engineering, I have studied psychology and follow the social science literature on safety. But I am sure that my primary education in engineering and mathematics affects my views of other fields. The following is what I interpret from what I have read about workplace safety and have found in my limited experiences working in this field in the past few years.

As seen in the previous historical narratives, before the industrial revolution, workers were expected to provide their own tools, to understand the risks associated with their trade, and to accept personal responsibility for their own safety. This attitude was justified partly by the fact that workers devoted their entire careers to the manufacturing of one or two products [Rogers, 1971]. They could, therefore, thoroughly understand their jobs and had control over how they performed their tasks.

The world has changed drastically since that time, although the same attitude about personal and workplace safety still exists today in many companies. The workers are often blamed for any accidents although they no longer have the freedom to control their own work situations: for the most part, workers do not have total control over their tools, the way they do their jobs, and the environment in which they work. Because of these changes, responsibility for workplace safety has shifted, or should have shifted, to those creating the workplace environment and tools, i.e., from the employee to the employer. The recognition of this shift did not come easily, and still is not totally accepted everywhere.

In some cases, government has had to step in to protect workers. As an example, in the nineteenth century, one of the main causes of injury and death to railroad workers occurred when workers were required to couple and decouple train cars. In the seven years between 1888 and 1894, 16,000 railroad workers were killed in coupling accidents and 170,000 were crippled. Managers claimed that such accidents were due only to worker error and negligence, and therefore nothing could be done aside from telling workers to be more careful. Because of the huge numbers of accidents, the government finally stepped in and required that automatic couplers be installed. As a result, fatalities and injuries dropped sharply. Three years after Congress acted on the problem, the June 1896 issue of Scientific American contained the following:

> *Few battles in history show so ghastly a fatality. A large percentage of these deaths were caused by the use of imperfect equipment by the railroad companies; twenty years ago it was practically demonstrated that cars could be automatically coupled, and that it was no longer necessary for a railroad employee to imperil his life by stepping between two cars about to be connected. In response to appeals from all over, the U.S. Congress passed the Safety Appliance Act in March 1893. It has or will cost the railroads $50,000,000 to fully comply with the provisions of the law. Such progress has already been made that the death rate has dropped by 35 per cent.*

By focusing on the operators in Safety-II, whether on what they do right or how they adapt in the face of hazards or even what they do wrong, operators become responsible for safety rather than focusing on the design of the system as a whole and the role of the operator within it. In most industrialized countries today, employers are expected to provide a safe working environment and the necessary tools and equipment to maintain that environment.

In Prof. Hollnagel's discussion of workplace history, I was puzzled by the extensive discussion of Taylorism or Scientific Management. Scientific Management principles were introduced to improve productivity and efficiency on assembly lines, not safety. It did have some unfortunate consequences in terms of new types of work place injuries, particularly due to repetitive motions. But otherwise, safety is unrelated to Taylorism. Prof. Hollnagel does admit on page 43(1) that Scientific Management does not consider safety, but claims that it "*had consequences for how adverse events were studied and for how safety could be improved.*" In my 40 years in this field, I have never seen any hint that Taylorism has

impacted how adverse events are studied or beliefs about how safety could be improved.[6] Prof. Hollnagel really needs a reference here as I cannot figure out how they are related. The sentence or two he provided seem to me to be a stretch and proof of nothing.

One of the important features of workplace safety that is not emphasized by Prof. Hollnagel has been its focus on behaviorism. Behavioral psychology is a theory of learning, advanced by Watson, Skinner and others, that emphasizes behaviors that are acquired through *conditioning*. Conditioning occurs through interaction with the environment. Strict behaviorists believe that any person can potentially be trained to perform any task, regardless of genetic background, personality traits, and internal thoughts (within the limits of their physical capabilities). It only required the right conditioning. Simply put, strict behaviorists believe that all behaviors are the result of experience.[7] Behaviorism was the dominant school of thought in psychology from about 1920 through the mid-1950s, when it began to be replaced by other theories. Instances of behavioristic thinking remain in workplace safety even today.

There are two types of conditioning, but the one most relevant to workplace safety is *operant conditioning*. Here learning is thought to occur through reinforcements and punishments. Basically, an association is made between a behavior and a consequence for that behavior. Most people have heard about Pavlov and his dogs. In the case of workplace safety, behaviors deemed to promote safety, such as wearing personal protective equipment (PPE) like hardhats or following specified procedures, are rewarded and those considered to be unsafe, such as not wearing PPE or acting recklessly, are punished. Thus, in workplace safety, *compliance* with specified procedures is emphasized, as Prof. Hollnagel notes. However, he does not associate this only with workplace safety and implies that it is also the standard approach to safety in product/system safety. As we will see, the role of procedures and compliance with procedures is much more nuanced in safety engineering and depends on the particular industry.

As described above, workplace safety and product/system safety had many overlaps historically, but at least 70 years ago they diverged and have taken totally different paths since that time with almost no interaction between the two. There are different educational programs, different certification requirements, different oversight agencies with very different approaches to preventing losses, different regulations, different management structures within an organization (EHS[8] groups usually oversee workplace safety while product/system safety is part of product development and engineering), different conferences and textbooks/journals, and even different models of causality or beliefs about how and why accidents occur (which is discussed later). Most important, they use very different practices to achieve safety goals. Basically they are two different fields with very little, if any, overlap.

I have seen a few instances where companies mixed the two (workplace and product safety) in one management structure, but usually that was when I was on the team investigating a serious system safety accident and the mixture was itself one of the major factors identified as being involved in the loss. An example is the Texas City Refinery explosion and the Deepwater Horizon blowout. In most industries, the training, practices, and management of these two fields are very different, and there is recognition that experts in one are not experts in the other and that mixing them may lead to neither being done well. The two groups within a company, in my experience, usually have little or no interactions nor any common practices.

Prof. Hollnagel's concept of Safety-I is much closer to workplace safety than safety engineering, as one can see by the emphasis in his book on Heinrich, Heinrich's Pyramid, and other workplace safety topics with almost no mention of product/system safety examples or most of the practices. So let's look at product/system safety as it has been practiced for the past 70 years.

---

[6] The only reason I know about Taylorism is that I have a graduate degree in management, and it was taught as an old and obsolete management theory from a hundred years ago.

[7] For a basic introduction to behaviorism, see https://www.verywellmind.com/behavioral-psychology-4157183

[8] Environmental Health and Safety

An Engineer's View of Product/System Safety

To understand the technical history of safety, a little terminology is needed, particularly the concept of a hazard, which surprisingly is barely mentioned in Prof. Hollnagel's books although it has been the basis for almost all of product/system safety for at least a century.

Engineers focus on eliminating, preventing, and responding to *hazards* rather than focusing on the behavior or actions of humans or even the accidents themselves (as Prof. Hollnagel does). Hazards are, informally, states of the system that can lead to losses, not the losses themselves. Losses are usually labeled as accidents, mishaps, or incidents.

Why was the concept of hazard introduced? An accident may involve conditions over which the engineer or designer of the system has no control. Consider a chemical plant. Accidents can happen when chemicals are inadvertently released from the plant, atmospheric conditions are such that the chemicals are transferred to a place where humans are present (instead of harmlessly dissipating into the atmosphere), and there are at the time humans or other relevant entities in the path of the chemicals. The only part of these events that the designer of the chemical plant has control over is the inadvertent release of the chemicals. That does not involve a loss itself but is a state that can lead to a loss under worst case conditions. That state is called a *hazard*. Examples for automobiles are violation of minimum distance with the car ahead of you or change into a lane where there is another car. In the first case, braking distance may depend on visibility and road conditions. In the second, the other driver in the other lane may be able to avoid you, but you don't have any real control over their behavior and the state of their car. Therefore, the focus is on preventing hazards, which we *can* potentially control.

Product/system safety encompasses a wide variety of practices. To try to reduce it to one simple stereotypical approach would be misleading: there is no monolithic approach to our attempts to reduce losses. The approach to safety varies with each industry and sometimes each company in an industry. There are some differences between countries, but for the most part, the few geographic differences are swamped by the differences between industries. Those differences usually arise due to the different nature of the hazards in those industries, but they also arise from different history and traditions in an industry and their general safety culture. I have not, however, come across *any* industry that approaches safety in the way described by Prof. Hollnagel as Safety-I.

System safety or product safety is usually taught as part of a particular engineering discipline and not called out as a separate topic. Safety engineering is based on the same scientific and mathematical foundations as other aspects of engineering. The procedures used to improve safety are those used in engineering, primarily modeling and analysis, in this case applied to hazardous system states. The goal is to identify hazards and then to design, manufacture, operate, and manage systems such that the identified hazards are eliminated or mitigated to prevent losses. In addition, within engineering, the term "safety" is usually broadly defined to include not just human injury (as in workplace safety) but also damage to physical equipment—such as an aircraft or spacecraft—and even mission losses in situations where human health or even equipment losses may not be a concern but the mission is critical.

I wrote a 687-page book [Leveson, 1995], that describes safety engineering practices up to the time it was published. More changes have occurred in the 25 years since that time. Clearly I cannot go into detail here, so I will provide an overview to counter Prof. Hollnagel's claim that the focus of safety engineering is reactive and on after-the fact accident investigation. My counter argument to his additional mischaracterization that the focus is on "what goes wrong" is covered later in this paper.

Activities Common among Different Industries

There are some activities common to most safety engineering efforts, although they may be performed very differently in different industries. These include:

Hazard Analysis: Hazard analysis is used to identify hazardous states during system design and development and to identify how the system can get into the hazardous states. Many different techniques have been developed for performing hazard analysis but all start from the hazards identified by the stakeholders and, using basic scientific knowledge about the technology involved along with what is known or assumed about how accidents occur in that industry, generate the scenarios or behaviors of the system components that can lead to the hazardous states.

Design for Safety: The goal here is to use the information obtained from the hazard analysis to engineer or "design out" accidents before they occur, that is, to prevent hazards. This is accomplished, in general, by design processes that are unique to various industries and the characteristics of the hazards in those industries. The highest priority is to create designs that cannot get into hazardous states. If that goal is not possible or practical, then (in priority order) an attempt is made to design to prevent hazardous states from occurring, to reduce their occurrence, or to respond if they do occur and move the system back into a safe or safer state in order to prevent or reduce losses.

The primary goal is to *eliminate hazards* from the original design. Hazards are eliminated either by eliminating the hazardous state from system operation (create an intrinsically safe design) or by eliminating the negative consequences (losses) associated with that state—if a state cannot lead to any potential losses, it is not a hazard. Of course, philosophically, almost nothing is impossible. Theoretically, you could be hit by a meteorite while reading this paper. But from a practical engineering standpoint, the occurrence of some physical conditions or events is so remote that their consideration is not reasonable. Hazard elimination may involve substituting a non-hazardous or less hazardous material for another one. Examples include substituting nonflammable materials for combustible ones or nontoxins for toxins.

There are almost always tradeoffs involved. Some hazards simply cannot be eliminated without, at the same time, resulting in a system that does not satisfy its goals. Taking more risks in a spacecraft design may be more justifiable than in a commercial airliner. All losses do not have the same priority.

If hazards cannot be eliminated, then an attempt is made to reduce their occurrence. Interlocks, increased reliability (if component failures are a cause identified in the hazard analysis), building in safety margins, etc. Examples include pressure release valves being designed with interlocks to prevent all the valves from being shut off simultaneously, devices to disable a car's ignition unless the automatic shift is in "*park*," the freeze plug in an automobile engine cooling system whose expansion will force the plug out rather than crack the cylinder if the water in the block freezes. Similarly, to protect against excessive heat, a fusible plug in a boiler becomes exposed when the water level drops below a predetermined level and therefore the heat is not directed away from the plug, which then melts. The opening permits the steam to escape, reduces the pressure in the boiler, and eliminates the possibility of an explosion. Safety margins involve designing a component to withstand greater stresses than are anticipated to occur. Redundancy can be used to introduce duplication to improve reliability, although there are many disadvantages and limitations of this approach such as common mode and common cause failures of redundant components. In addition, redundancy is not useful for software components or for component design errors.

Because it is often impossible to ensure that the design prevents the occurrence of all hazards, there is usually also an attempt to design to control hazards before damage occurs if, despite efforts to eliminate or reduce them, they do occur. For example, even with the use of a relief valve to maintain pressure below a particular level, a boiler may have a defect that allows it to burst at a pressure less than the relief valve setting. To deal with this, building codes may limit the steam boiler pressure that can be used in densely populated areas. Protection designed into the system to deal with preventing or limiting damage in case the hazard occurs may include designing to limit exposure of the hazard, isolation and containment, protection systems and fail-safe design.

Finally, the last resort is to design to reduce potential damage in the case of an accident, such as alarms and warning systems, contingency planning, providing escape routes (e.g., lifeboats, fire escapes, and community evacuation), or designs that limit damage (e.g., blowout panels, collapsible steering columns on cars, or shear pins on motor-driven equipment).

Human factors engineering and human-centered design: In human factors engineering, psychological concepts are applied to engineering designs to prevent human errors and provide human operators with the ability to safely control the system. More focused human-centered design concepts started to be developed in the 1980s and were first applied in the aviation community. In this approach to engineering design, the role of humans in the control of systems is the focus from the beginning of engineered system concept development.

Operations: Systems must not only be designed to be safe, but they must also be operated safely. Operational safety involves considerations for operability in the original design and for managing operations to ensure proper training of operators, identifying and handling leading and lagging indicators of risk, management of change procedures, maintenance procedures, etc. Data collection and analysis during operations has played an important role in improving design and operational safety.

Management and Policy: Emphasis on the design of Safety Management Systems is a relatively recent emphasis in system safety engineering, dating back to the middle of the last century. Claims are made in Prof. Hollnagel's book that these started in the 1990s, but that is not true. These efforts involve creating effective safety cultures, information systems, and safety management structures.

Accident investigation and analysis: Every industry investigates accidents, but it is usually a small part of the safety engineering effort.

Regulation and Licensing: Regulation may involve rules enforced by an oversight agency, voluntary standards, or certification/licensing of new systems. Regulation usually involves some type of approval of new systems before they are allowed to be used. It also almost always includes oversight into the operation of the systems to ensure that assumptions about operation (such as maintenance assumptions) made during analysis, design, and certification of the system hold in the operational environment and that changes over time are not leading to increasing levels of risk. If such dangerous conditions are caught in time, accidents can be prevented. Examples of ways that oversight agencies collect information during operations include licensee event reports in nuclear power plants, aviation safety reporting systems, and auditing of airline and airport operations.

In summary, as described above, the primary emphasis in almost all product/system safety fields is on the engineering of systems to prevent accidents through design and analysis activities during product development and by controls and feedback during operations. Most every industry and company also investigates accidents—it would be irresponsible not to do so: The information obtained is invaluable in improving prevention activities and our basic engineering knowledge. No industry that I know of, however, depends on investigation alone or even gives it priority over proactive prevention.

As noted, technology and the basic hazards differ in different industries so it is not possible to describe one approach that is used in all of them as Prof. Hollnagel does in the description of Safety-I. There are many more differences than similarities. But a few examples are given to provide some understanding of how safety engineering has been done for the past 70-100 years.

Commercial Aviation

Aircraft accident rates are very low. But this was not always the case. Aircraft accident rates peaked in the 1970s and have been driven down in the last 50 years even with a dramatic increase in the number of flights. Aircraft accident rates were so high in the 1950 and 1960s that only 20% of Americans were willing to fly [Leveson, 1995]. Bill Boeing realized that if he was to build a successful industry, accident rates needed to improve dramatically

This effort has been remarkably successful. That success was a result of doing many things, not simply investigating accidents.

Aircraft operate in widely differing and very challenging environments. At the same time, high risk is simply unacceptable. The approach used today for safety in commercial aircraft recognizes this reality.

Conservatism and a slow pace in design change is emphasized with occasional major technological developments, such as, the jet engine, the introduction of computers into control, and structures made of composites, that introduced new hazards but also eliminated others and, on balance, usually reduced accident rates or at least particular types of accidents. Except for the disruptive effects of these new technologies, the hazards have not changed significantly. Accidents are assumed to be caused by failures of the aircraft components.

When a new aircraft design is created, it is analyzed to identify failure scenarios that could lead to a hazardous state and then such scenarios are used by the design engineers as described above. Because there is often no safe state to move the system into (as in nuclear power where the nuclear reaction can be shut down), focus has been on using fault tolerance and high-integrity design and components. The design techniques used include design integrity and quality, introducing margins or factors of safety into the basic design, redundancy or backup systems, error tolerance, failure warning and indications to human operators as well as information about the current state of the aircraft or spacecraft, isolation of components or subsystems so that failure in one will not affect others, designed failure paths, damage tolerance, failure containment, redundancy and backup systems, design integrity and quality, flight crew training and procedures to prevent common errors and for use in situations where there is not enough time or information for human operators to figure out what to do themselves, and error tolerance (protection against common human errors in the design itself). Human factors engineering is an important part of the design of aircraft, particularly the design of cockpits and procedures.

Beyond the general design procedures, there are also specific airworthiness criteria including the design of seat belts, oxygen masks and systems in case of rapid depressurization, escape and evacuation procedures and equipment, life vests and rafts, etc. Many of these designs were improved after accidents showed some of the assumptions used in the design were flawed.

As underlying scientific knowledge about flight, materials, and propulsion was developed, this knowledge was used to design safety into aircraft and to prevent physical failures and design errors.

In addition to basic design techniques, operability and safety during operations are major considerations in the design process such as creating practical functional check procedures that can be used to determine a component's condition and its ability to perform its intended function. Such design for operability and maintainability includes periodic inspections, for example of the aircraft structure and engine rotor disks for fatigue damage.

But the design of the aircraft itself is a small part of the reason for low accident rates in this industry. An important factor is government and international regulatory agencies as well as independent industry and user associations, such as the airline pilots associations, the Flight Safety Foundation, the NASA aviation safety program, airline passenger associations, etc. Nearly all aspects of operations are regulated and audited. Programs for reporting near misses have been enormously successful in getting feedback before accidents occur. This topic is discussed more later.

Investigation of accidents has been an important component of reducing accidents in commercial aviation, particularly now that accident rates have been driven down so low. Investigations are remarkably thorough with information gathered about every component of the accidents. New components have been added to aircraft in order to assist in investigations, such as black box recorders, cockpit voice recorders, and other types of aircraft instrumentation.

There was a time, very early in the development of aviation, when "fly-fix-fly" *was* a common approach. But that was before the development of aerodynamics and the science behind flight. Before much is known scientifically about a new technology, the only way to learn is through experimentation.

But it is not the primary approach today except for very new concepts. Such experimentation is not done while endangering the public's lives, but instead may be done using test pilots and various types of experimental rigs, such as wind tunnels.

The effort put into investigation of accidents today is a result of the success of all the other things to increase safety in this industry. Accidents now occur because of changes in the design, operations, and environment of aircraft and when the assumptions underlying the original design and even oversight activities are no longer effective. We can try to anticipate that the assumptions are changing, but when the change is gradual and not planned, we are often left to gather this information after a tragedy. Again, it would be gross irresponsibility not to investigate accidents as thoroughly as possible to prevent repetition. It would also be gross irresponsibility to base aviation safety on investigation of accidents.

Nuclear Power

Nuclear power has in its short history developed a very different approach to safety, at least with respect to design, than that used in aviation. In addition to the safety engineering problems of other new and potentially dangerous technologies, the nuclear power industry has a relatively unique problem with public relations and has had to put a great deal of energy into convincing government regulators and the public that the plants are safe. This requirement, in turn, has resulted in a greater emphasis on probabilistic methods of risk assessment: the time required for empirical evaluation and measurement of safety is orders of magnitude greater than the modern pace of technological development. Another result is that nuclear power engineering has, until recently, used a few designs for which a lot of past experience can be accumulated and has been very conservative about introducing new technology, such as digital systems. That conservatism is giving way to greatly increased use of digital instrumentation and control.

In an effort to promote the development of nuclear power and also because the hazards were only partly understood, the industry was exempted from the requirements of full third-party insurance in some countries, such as the Price-Anderson Act in the United States [Thomson, 1987]. Instead, government regulation and certification were substituted as a means of enforcing safety practices in the industry.

The first nuclear power plant designs and sizes were also limited, although this has changed somewhat over the years as confidence in the designs and protection mechanisms has grown. The primary approach to safety in the nuclear industry is *defense in depth,* which is not the primary approach used in other industries. Defense-in-depth includes [National Nuclear Energy Policy Group, 1977]:

- A succession of barriers to a propagation of malfunctions, including the fuel itself; the cladding in which the fuel is contained; the closed reactor coolant system; the reactor vessel; any additional biological shield (such as concrete); and the containment building (usually including elaborate spray and filter systems).
- Primary engineered safety features to prevent any adverse consequences in the event of malfunction.
- Careful design and construction, involving review and licensing in many stages.
- Training and licensing of operating personnel.
- Assurance that ultimate safety does not depend on correct personnel conduct in case of an accident.
- Secondary safety measures designed to mitigate the consequences of conceivable accidents.
- Oversight of operations and licensee event reports.

Licensing is based on identification of the hazards, design to control these hazards under normal circumstances, and backup or shutdown systems that function in abnormal circumstances to further

control the hazards. Most of the emphasis, with respect to safety, is placed on the shutdown system that is brought into operation if a hazard occurs. The backup system designs are based upon the use of multiple, independent barriers, a high degree of single element integrity for passive features, and the provision that no single failure of any active component will disable any barrier. Note that siting nuclear power plants in remote locations is a form of barrier in which the separation is enforced by isolation. Because of the difficulty in isolating plants, emergency planning has gotten more attention since the Three Mile Island accident.

The dependence on defense-in-depth using barriers and reversion to a "safe state" (which does not usually exist in aviation and other fields) is unique to nuclear power. While specific details may differ between countries, redundancy and high integrity parts are commonly used and engineers have emphasized high component reliability and protection against component failure, which can lead to failed barriers and inability to reach a safe state.

Because many people outside the United States are most familiar with nuclear power as the most common highly safety-critical industry in their region, there have sometimes been assumptions that this is the only approach to safety engineering, but the design and general safety engineering approaches used in aerospace and defense are very different.

With the nuclear power, defense-in-depth approach to safety, an accident requires a disturbance in the process, a protection system that fails, and inadequate or failed physical barriers. These events are assumed to be statistically independent because of differences in their underlying physical principles: A very low calculated probability of an accident can be obtained as a result of this independence assumption [Rasmussen, 1987].

Recovery after a problem has occurred depends on the reliability and availability of the shutdown systems and the physical barriers. A major emphasis in building such systems, then, is on how to increase this reliability, usually through redundancy of some sort.

Certification of nuclear power plants has emphasized probabilistic risk assessment that includes estimating the reliability of the barriers and protection systems. Because empirical measurement of probability is not practical for nuclear power designs, most approaches build a model of accident events and construct an analytical risk assessment based upon the model. Such models include the events leading up to an uncontrolled release as well as data on factors relating to the potential consequences of the release, such as weather conditions and population density.

In addition to probabilistic risk assessment in certification of nuclear plants, other features used in nuclear power safety include incident and error reporting, an emphasis on creating a strong safety culture, external oversight and regulation, and yes, in-depth accident investigation.

Chemical Industry

The process and materials in the chemical (process) industry are inherently hazardous, and the approach to safety was driven by insurance needs. Thus, the term usually used in the industry, *loss prevention*, reflects these origins.

In chemical and petrochemicals, the three major hazards are fire, explosion, and toxic release. Of these three, fire is the most common, explosions are less common but cause greater losses, and toxic release is relatively rare but has been the cause of the largest losses. Because loss of containment is a precursor for all of these hazards, much of the emphasis in loss prevention is, like the nuclear industry (which is usually considered part of the process industry), on avoiding the escape of explosive or toxic materials through leaks, ruptures, explosions, and so on.

The three major hazards related to the chemical industry have remained virtually unchanged in their nature for decades. Design and operating procedures to eliminate or control these hazards have evolved and been incorporated into codes and standards produced by the chemical industry societies and

others. This approach sufficed before World War II because the industry operated on a relatively small scale and development was slow enough to learn by experience. After World War II, however, the chemical and petrochemical industries began to grow in complexity, size, and new technology at a tremendous rate. The potential for major hazards grew at a corresponding rate. The operation of chemical plants also has increased in difficulty, and startup and shutdown have especially become complex and dangerous.

The effect of the changes has been to increase the consequences of accidents, to increase environmental concerns, such as pollution and noise, to make the control of hazards more difficult, and to reduce the opportunity to learn by trial and error. At the same time, the social context has been changing. In the past, safety efforts in the process industries were primarily voluntary and based on self-interest and economic considerations. However, pollution has become of increasing concern to the public and to government. Major accidents have drawn enormous publicity and generated political pressure for legislation to prevent similar accidents in the future. Most of this legislation requires a qualitative hazard analysis, including identification of the most serious hazards and their contributing factors and modeling of the most significant accident potentials [Rouhiainen, 1990].

These factors led the industry to increase its proactive efforts to analyze potential hazards more carefully and to reduce emissions and noise. The accidental release of MIC (methyl isocyanate) at Bhopal, however, demonstrated that technical hazard analysis by itself is not enough and that management practices may be even more important.

Applying hazard analysis in the chemical industry has special complications compared to other industries, which makes the modeling of causal sequences especially difficult [Suokas, 1988]. Although the industry does use some standard reliability and hazard analysis techniques, the unique aspects of the industry have led to the development of industry-specific techniques. For example, the hazardous features of many chemicals are well understood and have been catalogued in indexes that are used in the design and operations of plants.

A hazard analysis technique developed for and used primarily in the chemical and petrochemical industries is called Hazards and Operability Analysis (HAZOP). This technique is a systematic approach for examining each item in a plant to determine the causes and consequences of deviations from normal plant operating conditions. The information about hazards obtained in this study is used to make changes in design, operating, and maintenance procedures. HAZOP dates from the mid-1960s. Newer hazard analysis techniques (such as STPA) can be performed earlier (before the design is complete) and the results used to build simpler, cheaper, and safer plants by avoiding the use of hazardous materials, using less of them, or using them at lower temperatures or pressures.

Accident investigation is of course done if losses occur, but proactive analysis and design efforts are much more heavily emphasized.

Defense and “System Safety”

After World War II, the United States Department of Defense created a unique approach to safety that they called System Safety [Miller, 1985]. The U.S. started to build nuclear weapons, launch on warning systems and early warning systems, intercontinental ballistic missiles, etc. The consequences of accidents in these systems were so enormous that safety had a high priority. Some defense systems, such as atomic weapons, potentially have such catastrophic consequences that they must not be allowed to behave in an unsafe fashion. The primary approach in such systems is to identify hazards and then try to eliminate them in the system design. Accidents to investigate are few and far between. Near misses may occur but the emphasis is not on after-the-fact investigation but instead on building systems to be safe from the beginning.

Many of the features of System Safety, because of the Cold War, were not known outside a small group of engineers and companies that worked on these very secret systems. The approach is no longer

secret, but it is largely unknown outside the U.S. or even by many safety practitioners in other industries within the U.S.

Much of the early development of System Safety as a separate discipline in the defense industry began with flight engineers after World War II. The Air Force had long had problems with too many aircraft accidents. For example, from 1952 to 1966, it lost 7715 aircraft, in which 8547 people, including 3822 pilots, were killed [Hammer, 1980]. Most of these accidents were blamed on pilots. Many industry aeronautical engineers, however, did not believe the cause was so simple. They argued that safety must be designed and built into aircraft just as are performance, stability, and structural integrity [Miller, 1985; Stieglitz, 1948].

Seminars were conducted by the Flight Safety Foundation, headed by Jerome Lederer (who would later head the NASA Apollo safety program), that brought together engineering, operations, and management personnel. It was in 1954, at one of these seminars, that the term "system safety" may have first been used—in a paper by one of the aviation safety pioneers, C.O. Miller. Around the same time, the Air Force began holding symposiums that fostered a professional approach to safety in propulsion, electrical, flight control, and other aircraft subsystems, but they did not at first treat safety as a system problem.

When the Air Force began to develop intercontinental ballistic missiles (ICBMs), there were no pilots to blame for accidents, yet the liquid-propellant missiles blew up frequently and with devastating results. The missiles used cryogenic liquids with temperatures down to –320°F. and pressurized gases at 6000 pounds per square inch, and the potential safety problems could not be ignored—the highly toxic and reactive propellants were sometimes more lethal than the poison gases used in World War II, more violently destructive than many explosives, and more corrosive than most materials used in industrial processes [Hammer, 1980]. The Department of Defense and the Atomic Energy Commission (AEC) were also facing the problems of building and handling nuclear weapons and finding it necessary to establish rigid controls and requirements on nuclear materials and weapon design.

In that same period, the AEC (and later the NRC or Nuclear Regulatory Commission) was engaged in a public debate about the safety of nuclear power. Similarly, civil aviation was attempting to reduce accidents in order to convince a skeptical public to fly, and the chemical industry was coping with larger plants and increasingly lethal chemicals. These parallel activities resulted in different approaches to handing safety issues, as described above.

System Safety itself arose out of the ballistic missile programs. In the fifties, when the Atlas and Titan ICBMs were being developed, intense political pressure was focused on building a nuclear warhead with delivery capability as a deterrent to nuclear war. In an attempt to shorten the time between initial concept definition and operational status, a concurrent engineering approach was developed and adopted. In this approach, the missiles and facilities in which they were to be maintained ready for launch were built at the same time that tests of the missiles and training of personnel were proceeding. The Air Force recognized that this approach would lead to many modifications and retrofits that would cost more money, but with nuclear war viewed as the alternative, they concluded that additional money was a cheap way to buy time. A tremendous effort was exerted to make the concurrent approach work [Rogers, 1971].

On these first missile projects, system safety was not identified and assigned as a specific responsibility. Instead, as was usual at the time, each designer, manager, and engineer was assigned responsibility for safety. These projects, however, involved advanced technology and much greater complexity than had previously been attempted, and the drawbacks of the standard approach to safety became clear when many interface problems went unnoticed until it was too late.

Within 18 months after the fleet of 71 Atlas F missiles became operational, four blew up in their silos during operational testing. The missiles also had an extremely low launch success rate. Launch failures were caused by inadequate requirements, single point failures, construction errors, bypass of

procedural steps, and by management decisions to continue in spite of contrary indications because of schedule pressures [Air Force Space Division, 1987].

Not only were the losses themselves costly, but the resulting investigations detected serious deficiencies in the systems that would require extensive modifications to correct. In fact, the cost of the modifications would have been so high that a decision was made to retire the entire weapon system and accelerate deployment of the Minuteman missile system. Thus, a major weapon system, originally designed to be used for a minimum of ten years, was in service for less than two years primarily because of safety deficiencies [Rogers, 1971].

When the early aerospace accidents were investigated, it became apparent that the causes of a large percentage of them could be traced to deficiencies in design, operations, and management. The previous *fly-fix-fly* approach was clearly not adequate. In this approach, investigations were conducted to resurrect the causes of accidents, action was taken to prevent or minimize the recurrence of accidents with the same cause, and eventually these preventive actions were incorporated into standards, codes of practice, and regulations. Although the fly-fix-fly approach was effective in reducing the repetition of accidents with identical causes, it became clear to the Department of Defense (DoD), and later others, that it was too costly and, in the case of nuclear weapons, unacceptable. This recognition led to the adoption of System Safety approaches 60–70 years ago to try to prevent accidents before they occur the first time.

The Army soon adopted System Safety programs developed by the Air Force because of the many personnel it was losing in helicopter accidents, and the Navy followed suit. In 1966, the DoD issued a single directive requiring System Safety programs on all development or modification contracts.

At first there were few techniques that could be used on these complex systems. But, step-by-step, scientific, technical, and management techniques were developed or adapted from other activities. Contractors were required to establish and maintain a System Safety program that is planned and integrated into all phases of system development, production, and operation. A formal plan was required that ensured that

1. Safety, consistent with mission requirements, is designed into the system.
2. Hazards associated with the system, each of the subsystems, and the equipment are identified and evaluated and eliminated or controlled to an acceptable level.
3. Control over hazards that cannot be eliminated is established to protect personnel, equipment, and property.
4. Minimum risk is involved in the acceptance and use of new materials and new production and testing techniques.
5. Retrofit actions required to improve safety are minimized through the timely inclusion of safety engineering activities during the acquisition of a system.
6. The historical safety data generated by similar programs are considered and used where appropriate.

To summarize all of this, the primary concern of System Safety is the management of hazards: their identification, evaluation, elimination, and control through analysis, design, and management procedures. System Safety activities start in the earliest concept development stages of a project and continue through design, production, testing, operational use, and disposal. The primary emphasis is on the early identification and classification of hazards so that corrective action can be taken to eliminate or minimize their impact before final design decisions are made.

To understand the unique aspects of the System Safety approach and differentiate it from the other approaches to safety developed in parallel but independently for such industries as civil aviation and nuclear power, a few basic concepts can be identified:

- *System Safety emphasizes building in safety, not adding it on to a completed design or trying to assure it after the design is complete*.

From 70 to 90% of the design decisions that affect safety will be made in the concept development project phase [140]. The degree to which it is economically feasible to eliminate a hazard rather than to control it depends on the stage in system development at which the hazard is identified and considered. Early integration of safety considerations into the system development process allows maximum safety with minimal negative impacts. The alternative is to design the system or product, identify the hazards, and then add on protective equipment to control the hazards when they occur—which usually is more expensive and less effective. Waiting until operations and then expecting human operators to deal with hazards—perhaps by assuming they can be flexible and adaptable, as in Safety-II—is the most dangerous approach.

- *System safety deals with systems as a whole rather than with subsystems or components.*
  Safety is a system property, not a component property.
- *System safety takes a larger view of hazards than just failures.*
  Serious accidents can occur while system components are all functioning exactly as specified—that is, without failure. In addition, the engineering approaches to preventing failures (increasing reliability) and preventing hazards (increasing safety) are different and sometimes conflict.
- *System safety emphasizes analysis rather than past experience and standards.*
  Standards and codes of practice incorporate experience and knowledge about how to reduce hazards, usually accumulated over long periods of time and resulting from previous mistakes. While such standards and learning from experience are essential in all aspects of engineering, including safety, the pace of change today does not allow for such experience to accumulate and for proven designs to be used. System safety analysis attempts to anticipate and prevent accidents and near-misses before they occur.
- *System safety emphasizes qualitative rather than quantitative approaches.*
  System safety places major emphasis on identifying hazards as early as possible in the design stage and then designing to eliminate or control those hazards. In these early stages, quantitative information usually does not exist. And our technology and innovations are proceeding so fast that historical information may not exist nor be useful. The accuracy of quantitative analyses is also questionable. The majority of factors in accidents cannot be evaluated in numerical terms, and those that can will often receive undue weighting in decisions based on absolute measures.

  In addition, quantitative evaluations usually are based on unrealistic assumptions that are often unstated, such as that accidents are caused by failures, failures are random, testing is perfect, failures and errors are statistically independent, and the system is designed, constructed, operated, maintained, and managed according to good engineering standards. Some components of high technology systems may be new or may not have been produced and used in sufficient quantity to provide an accurate probabilistic history of failure. Surprisingly few scientific experiments (given the length of the time they have been used) have been performed to determine the accuracy of probabilistic risk assessment, but the results of the few that have been done have not been encouraging.
- *System safety recognizes the importance of tradeoffs and conflicts in system design.*
  Nothing is absolutely safe, and safety is not the only or usually even the primary goal in building systems. Most of the time, safety acts as a constraint on how the system goals (mission) may be achieved and on the possible system designs. Safety may conflict with other goals such as operational effectiveness, performance, ease of use, time, and cost. System safety techniques focus on providing information for decision making about risk management tradeoffs.
- *System safety is more than just system engineering.*
  System safety concerns extend beyond the traditional boundaries of engineering to include such things as political and social processes, the interests and attitudes of management, attitudes and

motivations of designs and operators, human factors and cognitive psychology, the effects of the legal system on accident investigations and free exchange of information, certification and licensing of critical employees and systems, and public sentiment [Lederer, 1986].

What is labeled Safety-III in this paper incorporates and extends this System Safety approach.

SUBSAFE: The U.S. Nuclear Submarine Safety Program

While nuclear submarines are engineered using standard System Safety engineering, a special program, called SUBSAFE, to deal with some aspects of nuclear submarine safety was created after the loss of the USS Thresher in 1963. Before that time, the U.S. lost one submarine (in peacetime) about every two to three years. After SUBSAFE was created, there has not been a loss of a submarine that was in the SUBSAFE program in the past 57 years. Only the highlights are described here. For a more detailed description, see Leveson [2012].

SUBSAFE focuses on one hazard: Inability of critical systems to control and recover from flooding (i.e., loss of hull integrity). The other hazards are handled using standard System Safety as described above.

As there are no accidents to investigate, this clearly is not a focus of the program. There are incidents, however, and they are investigated just as in every other industry. But, once, again, incident investigation is a very small part of the efforts. The emphasis instead is on establishing and maintaining a strong safety culture, sophisticated risk management without the use of untestable or unverifiable arguments based on probabilistic risk assessment, certification and performance audits, sophisticated safety management, education and training, specification and documentation, and continuous improvement.

Astronautics and Space

The space program was the second major application area after defense to apply the System Safety approach in a disciplined fashion. Until the *Apollo 204* fire in 1967 at Cape Kennedy, in which three astronauts were killed, NASA safety efforts had focused on workplace safety. The accident woke up NASA, and they developed policies and procedures that became models for civilian aerospace activities [Miller 1987]. Jerome Lederer, one of the pioneers in aviation safety, was hired to head manned space flight safety, and, later, all NASA safety efforts. Through his efforts, an extensive program of System Safety (as described above) was set up for space projects, much of it patterned after the Air Force and DoD programs. Many of the same engineers and companies that had established formal System Safety programs for DoD contracts also were involved in space programs, and the technology and management activities were transferred to these new applications.

More recently, particularly after the Columbia Space Shuttle loss, NASA has hired safety engineers from the nuclear power community and much of the System Safety standards and approaches were replaced with the probabilistic risk assessment approach used in nuclear power.

Healthcare/Hospital Safety

Healthcare is, in many ways, closer to a service than a product. Much of safety in healthcare involves the treatment of patients, not the engineering of medical equipment, although engineered medical devices are playing an increasing role in healthcare. There are also special factors such as new hazards and diseases appearing all the time.

Workplace (hospital) safety and patient safety may be more closely intertwined than in other industries, although there are still significant differences. While it is normal for goals and hazards to conflict, healthcare is somewhat unique in that many of the hazards (and thus constraints) conflict. That

often makes finding acceptable solutions to safety problems more difficult. For example, consider the following hazards:

1. Patients get a treatment that negatively impacts their health
2. Patients do not get the treatment they need, which results in death or serious negative health impacts.

There are times when taking extraordinary risks may be justified in this context where either choice may involve serious losses. At the same time, much effort is spent in preventing unnecessary losses.

Other unique factors in healthcare are that it is information-intensive and human-intensive, complex, imprecise, interdisciplinary and constantly changing. Drucker has said that "*healthcare institutions are complex, barely manageable places […] Large healthcare institutions may be the most complex organizations in human history*" [Drucker, 2002]

An engineering approach to patient safety has never been emphasized. Patients are not "engineered," and biological and health information may be only partially known. But there are lots of engineered/designed components and processes in hospitals and healthcare including electronic health records; medical devices; pharmacy procedures; and special procedures used to eliminate or control known hazards such as patient handoffs, communication protocols, checklists, and identification devices such as RFID. Procedures such as hand washing are used to control infections, while other interventions are used to reduce wrong site surgery, etc. In recent years, there has been a tremendous increase in technology in healthcare. There are also management structures in most hospitals to manage quality and safety as well as a large number of patient safety organizations devoted to promoting safety by designing healthcare systems to prevent "adverse events," the common term for accidents or mishaps in healthcare.

Safety and resilience in medical care and hospitals clearly depends on more than the behavior of frontline workers including individual doctors, nurses, and technicians, which seems to be the emphasis in Prof. Hollnagel's Safety-II. Design of processes, facilities, and equipment are all important as well as training of personnel, the hospital culture, and management structures for safety and quality in hospitals.

Much focus in health care has been on after-the-fact investigation. There are, however, starting to be cultural changes such as instituting a "just culture" to change investigations from focusing on blame to learning how to improve the design of the system. The investigation techniques and practices largely prevalent in healthcare, in general, are not very sophisticated (compared to those used for engineered systems) and need to be improved.

One limitation of the current approaches in healthcare and hospital safety is that they lack a holistic, systems standpoint, and the attempts to improve safety, while sincere, have been largely piecemeal and disjointed. Like the proverbial blind men and the elephant, each is focused on one part of the "elephant" but misses the other parts. Too much emphasis and responsibility has been placed on individual people, such as doctors and nurses, who have only limited control over the operations of the hospital and the healthcare system as a whole. A more comprehensive approach could be achieved by using systems thinking and a systems approach to safety, labeled in this paper as Safety-III.

Summary

In summary, Prof. Hollnagel describes Safety I as primarily using fly-fix-fly approaches and emphasizing after-the-fact accident investigation. Throughout the book he characterizes Safety-I and current practice as almost totally reactive. The above summary of safety engineering as practiced in most fields, demonstrates this characterization is simply untrue. The closest is workplace safety, which as described above diverged from safety engineering a hundred years ago. While hospital safety has put effort into after-the-fact investigation of adverse events, there has always been more emphasis on preventing unnecessary deaths than simply investigating them. The influence of healthcare on Prof.

Hollnagel's view of safety can be seen in his frequent use of the term "adverse outcome," which is a term I have seen used only in healthcare. Safety engineering uses the terms accident, mishap, or loss.

As described above, the emphasis on designing safety into systems and products from the beginning is at least 100 years old. But the urgency increased after World War II as companies and industries found themselves faced with similar problems—products and manufacturing facilities that were becoming increasingly complex and increasingly dangerous. The use of after-the-fact accident investigation to eliminate the causes became uneconomical as modification, retrofit, and replacement costs soared and liability concerns increased.

In addition, many of today's complex systems require integrating parts designed and built by separate contractors or organizational entities. Even if each contractor or group takes steps to build quality into its own components, combining subsystems into a system introduces new paths to hazards that are not apparent when viewing the parts separately. It is becoming evident in many industries that designing safety into a plant or product from the beginning of development could reduce overall life-cycle costs. An updated System Safety approach as practiced in the most dangerous defense systems can achieve acceptable levels of safety even in extremely complex systems. In this paper, that approach is described as Safety-III. It is not new, but is an evolution of what has been done in product/system safety at least since World War II.

It's time now to look in more detail at Prof. Hollnagel's specific claims about Safety-I and Safety-II.

A Comparison of Safety-I, Safety-II, and Safety-III

Prof. Hollnagel's table summarizing Safety-I and Safety-II is reproduced in Table 1 below. It has been augmented with two additional columns (the shaded part): the fourth column describes what is actually done today (rather than Prof. Hollnagel's strawman description in the second column), and the fifth column describes Safety-III and what I believe is the most promising approach to take in the future.

The rest of the paper examines the statements made by Prof. Hollnagel about each of the five rows plus a few other incorrect claims that Prof. Hollnagel makes that are important to clarify. In addition, in his book, Prof. Hollnagel defines and uses engineering terms in a way that is different than they are defined and used in engineering. Where the different definitions are important in coming to different conclusions, the more standard definitions are provided and compared with Prof. Hollnagel's definitions.

Table 1: The first three columns are from Hollnagel []. The final two columns are added.

	Safety-I	Safety-II	Safety engineering today	Safety-III
Definition of Safety	As few things as possible go wrong	As many things as possible go right	Safety is usually defined as freedom from unacceptable losses as identified by the stakeholders, but may be defined in terms of acceptable risk or ALARP in some fields. The goal is to eliminate, mitigate, or control hazards, which are the states that can lead to these losses.	Safety is defined as freedom from unacceptable losses as identified by the system stakeholders. The goal is to eliminate, mitigate, or control hazards, which are the states that can lead to these losses.

Safety Management Principle	Reactive, respond when something happens, or is categorised as an unacceptable risk	Proactive, continuously trying to anticipate developments and events	Concentrates on preventing hazards and accidents but does learn from accidents, incidents, and audits of how system is performing.	Concentrates on preventing hazards and losses, but does learn from accidents, incidents, and audits of how system is performing.
Explanations of accidents	Accidents are caused by failures and malfunctions. The purpose of an investigation is to identify causes and contributory factors.	Things basically happen in the same way, regardless of the outcome. The purpose of an investigation is to understand how things usually go right as a basis for explaining how things go wrong.	Accidents are caused by linear chains of failure events. The purpose of investigation is to identify the chain of events and the root cause.	Accidents are caused by inadequate control over hazards. Linear causality is not assumed. There is no such thing as a root cause. The entire socio-technical system must be designed to prevent hazards; the goal of investigation is to identify why the safety control structure did not prevent the loss.
Attitude to the human factor	Humans are predominantly seen as a liability or a hazard.	Humans are seen as a resource necessary for system flexibility and resilience.	Humans are expected to prevent or respond to hazards and to be flexible and resourceful when they occur.	The system must be designed to allow humans to be flexible and resilient and to handle unexpected events.
Role of performance variability	Harmful, should be prevented as far as possible.	Inevitable but also useful. Should be monitored and managed.	The primary reason for performance variability is to enhance productivity and system requirements. Procedures are provided when response time and information is limited. Effort is put into providing appropriate controls and interfaces to allow operators to prevent or respond to hazards. Design so that when performance (of	Design the system so that performance variability is safe and conflicts between productivity, achieving system goals, and safety are eliminated or minimized. Design so that when performance (of operators, hardware, software, managers, etc.) varies outside safe boundaries, safety is still maintained.

			operators, hardware, software, managers, etc.) varies outside safe boundaries, safety is still maintained.	

Let's start with the first row, the basic definition of safety. Prof. Hollnagel is correct in his statement that the term is used differently in different fields and countries. The problem is that engineers do not deal in vague and undefined concepts like "goes right" and "goes wrong." Such terminology is not only meaningless, but it is misleading and is used by Prof. Hollnagel to derive untrue conclusions.

Definition of Safety

Even in engineering, there are different definitions of safety. A straightforward and inclusive one is simply that safety is the freedom from harm or other undesirable outcomes. In the U.S. defense department standard, MIL-STD-882 (in all its versions since 1969), safety is defined as "freedom from conditions that can cause death, injury, occupational illness, damage to or loss of equipment or property, or damage to the environment." The standard Oxford English Dictionary defines safety as "being protected from danger or harm." Other definitions may include different or additional types of losses such as a mission loss. There is no relativity in any of these definitions, such as depending on how much it would cost to reduce losses.

Complications arise with the introduction of the term "risk" for measuring safety. Most often, risk is defined as a combination of the severity and likelihood of an unwanted outcome. One problem that arises is that by defining something, such as risk, as only one way to measure it, any alternatives then become impossible including a non-probabilistic measurement. It would be better to define risk as an assessment of safety and then allow different approaches to performing that assessment.

Returning to the basic concept of safety, instead of being expressed as an absolute, safety is sometimes defined in terms of being "as low as possible." Prof. Hollnagel uses this phrase in his definition of Safety-I as the condition where "the number of adverse outcomes (accidents, incidents, or near misses) is as low *as possible*" [p. 183]. Elsewhere in the book, he defines Safety-I as "As few things *as possible* go wrong" [p. 147].

The problem occurs in determining how low is "possible." There are always tradeoffs made when multiple objectives are imposed on a system design. One might be able to reduce the number of negative outcomes but that would require sacrificing other goals or even increasing the number of other types of negative outcomes or it might require spending more money. What is possible may be determined by the perch from which one is observing the system: the person paying to lower risk, the potential victim of an accident, or the government regulator.

In addition, the definition of "as few things as possible go wrong" is completely undefined. What does "go wrong" mean? If I don't make as much money from the system as I hoped, then it has gone wrong for me but is the system unsafe? We will return later to Prof. Hollnagel's undefined and

misleading use of terms like “go wrong” and “go right” with respect to safety. It is one of the biggest problems in his argument for Safety-II and plays a role in most of his strawman arguments.

Another common approach is to talk about safety or risk being “as low as reasonably practicable” (abbreviated ALARP) rather than as low as possible. This definition is common in U.K. government standards but it also has serious philosophical difficulties. Once again, who decides what is reasonably practicable? The person reaping profits from a system may have a different view of what is reasonably practical than a person who may be injured by the system. Who decides how much money, time, or effort is reasonable to devote to making something safer? Does the definition change if oneself or one’s loved ones are likely to be injured or killed rather than a stranger? Is it misleading to say something is “safe” but there is still a significant potential for losses?

At this point, one gets into philosophical and ethical concerns that go beyond engineering and venture in the area that Alvin Weinberg labeled *trans-scientific*. Alvin Weinberg was an American nuclear physicist and administrator of Oak Ridge National Laboratory during and after the Manhattan Project. In a landmark paper, he wrote:

> *Many of the issues that lie between science and politics involve questions that can be stated in scientific terms but that are in principle beyond the proficiency of science to answer […]. I proposed the term “trans-scientific” for such questions. Although they are epistemologically speaking, questions of fact and can be stated in the language of science, they are unanswerable by science; they transcend science…In the current attempts to weigh the benefits of technology against its risks, the protagonists often ask for impossible scientific answers to questions that are trans-scientific* [Weinberg, 1972].

Weinberg contends that the debate on risks versus benefits would be more fruitful if we recognized those limits. That leaves the questions involving measurement like “how safe is safe enough” or “what is ‘reasonably practical’” or “how low is ‘as low as possible’” as ones that cannot be answered by scientists and engineers as they involve individual value systems.

So to enhance communication, I use the absolute definition that is common to the System Safety and U.S. defense industry: Safety is freedom from those conditions that can cause death, injury, occupational illness, environmental damage, or damage to or loss of equipment or property. That is, safety is freedom from losses defined as important to the system stakeholders. We can then measure how close we get to this ideal. Otherwise, everyone starts from a different definition of safety and communication is inhibited. As an analogy, would most people say that their bank accounts were secure if there are only two break-ins a month? If there was an “acceptable” number of break-ins a month? As low as possible break-ins a month? In these cases, we might all decide to keep our money under our mattresses.

Once the losses are defined, then the safety engineering process begins. *Hazard analysis*, as described briefly above, involves identifying the scenarios describing the conditions under which the system will be unsafe. If a car is traveling at 100 mph, the brakes may be ineffective in preventing an accident when the road is wet even if they are effective at lower speeds and under different conditions.

It is, of course, ideal if there are no conditions under which something will be unsafe, i.e., the hazards can be eliminated from the design. Achieving this goal may require unacceptable tradeoffs with the reason the system is being created. If elimination is not possible, the goal becomes to reduce the occurrence of hazards, but this is clearly less desirable than the first because it does not guarantee that they will not occur. A third goal is, if the hazards occur anyway, to prevent them from leading to an accident. If that is not possible, then the final recourse is to try to reduce the impact, but this is clearly less desirable than preventing losses even if the hazard occurs. More details in how to accomplish these design goals can be found in Leveson (1995).

This process of identifying hazards and then designing systems to eliminate, prevent, or minimize their impact is what safety engineering is all about. It is not investigating accidents, but of course we do that in order to learn where our previous efforts went wrong. We certainly do NOT just try to make the system operate safety by doing things "right." There is no "right" or "wrong," but only requirements specified for a particular system. Engineers of course focus on satisfying the requirements, but they also consider what happens in case "things go wrong" or, to use engineering language, hazards occur and what to do about them.

Let's get back to Prof. Hollnagel's argument.

> Page 94: *Since Safety-I is defined as the 'freedom' from (unacceptable) risks, the phenomenology is really this 'freedom', i.e., the fact that nothing happens.*

Yes, that is correct. A system is defined as safe when no losses occur, where the losses must be defined by the stakeholders. A system is secure if there are no break-ins. A system is reliable when there are no failures.

> Page 94: *However, as has been argued previously, Safety-I is in reality defined in terms of its opposite, namely the lack of safety.*

Why? First, engineers define safety in terms of hazards and losses. When there are no accidents or losses, the system has behaved safely. There is no need to play with words here.

> Page 94: *This means that we can say that a system is unsafe when something happens. We also tend to believe that a system is safe---or rather, not unsafe---when nothing happens, although this, strictly speaking, is not a valid logical conclusion.*

It is not clear who "we" is, but engineers do not deal in "beliefs." If no losses have occurred, then no losses have occurred and the system has behaved safely. Nobody would claim that a hazard that has not been eliminated cannot occur under some circumstances in the future and lead to a loss. We can only talk about behavior up to this time. We cannot predict the future (although some people try). A prediction about the future is part of the definition of "risk" but not of safety. Engineering does not define a system as "safe" or "unsafe." Those terms are undefined and not useful. Instead, defined terms such as hazards and losses are used.

Prof. Hollnagel's argument here is actually pretty strange because in his definition of Safety-II, he suggests that we should focus on behavior when nothing happens ("things go right") and assumes that this implies the behavior is safe. But even according to his definition above, we cannot assume that just because an accident has not occurred, it cannot in the future.

> Page 94-95: *This 'reverse' definition creates the interesting practical question, How to measure an increase in safety by counting how many fewer things go wrong?*

Engineers do not use absolute counts but instead talk about the occurrence or non-occurrence of particular hazards or losses. Safety is not a measurement or count. Measurement is used in the concept of "risk," but risk is not defined in terms of counting how many things have gone wrong, as implied in this quote. Measurement of risk is controversial and not everyone believes it is possible or feasible for most systems. But most safety engineers focus on the goal of eliminating or controlling hazards—not counting how many "fewer things" go wrong or even how many things go wrong. One inadvertent detonation of a nuclear bomb could be catastrophic and clearly unsafe. If there is a second one, we would not say that the single detonation is safer than if there were two instances. That would be a useless type of measurement. Measuring by using counting is not useful when discussing a trans-scientific property. In this context, "safer" is undefined.

A real problem does exist in trying to convince people that safety engineering efforts are effective and thus worth the investment when nothing happens. That is exactly why arguments should not be based on "how many things go wrong." Instead, arguments should be (and are) based on how the defined hazards have been eliminated or controlled in the design of the system, including the design of

operations. While people do tend to reduce their estimates of risk, incorrectly, when nothing goes "wrong," that does not at all deter us from working on making systems safer. Engineers define safety in terms of hazards and work to eliminate or prevent them. Note that security is also defined in terms of things we don't want, i.e., break-ins or intrusions. Note also that it is not possible to conclude that safety has increased if the number of things that go "right" (according to whom?) increases, as occurs in Prof. Hollnagel's definition of Safety-II.

Most of the problems in Prof. Hollnagel's arguments and conclusions arise because of his use of the vague and undefined terms "goes right" and "goes wrong" throughout his books instead of using standard safety and engineering terminology. So it is important to look at this.

"Goes Wrong" vs. "Goes Right"

The first row of Prof. Hollnagel's table is the following:

	Safety-I	Safety-II	System Safety Engineering	Safety-III
Definition of Safety	As few things as possible go wrong	As many things as possible go right	Safety is usually defined as freedom from unacceptable losses as identified by the stakeholders, but may be defined in terms of acceptable risk or ALARP in some fields. The goal is to eliminate, mitigate, or control hazards, which are the states that can lead to these losses.	Safety is defined as freedom from unacceptable losses as identified by the system stakeholders. The goal is to eliminate, mitigate, or control hazards, which are the states that can lead to these losses.

Safety-I, according to Prof. Hollnagel, defines safety as "as few things as possible go wrong" while his Safety-II defines it as "as many things as possible go right." Safety engineering uses neither of these terms. First, I don't know what "things" are.

Also, as an engineer and safety expert, I am unable to interpret the terms "goes wrong" and "goes right" even after searching Prof. Hollnagel's book to determine what he might mean by these phrases. I could find almost no examples and no formal definitions. This is just not the way that engineers or safety experts talk. The Venn diagram in Figure 2 shows the definitions used in engineering. "Goes right" and "goes wrong" and "things" would never be used because they are undefined and ambiguous.

The green circle in Figure 2 would presumably include Prof. Hollnagel's "things go right." The red and blue circles would include "things go wrong." Note, however, that a failure (either of a component or even the system) is not the same as an accident, although there may be some overlap. So either "going wrong" does not involve failures where there are no losses or it includes failures that include losses and thus includes things other than safety.

The real problem, however, is all the other space inside the black circle representing all of the other possible system behavior not included in the green, blue, and red circles. Is that "goes right" or "goes wrong"? Prof. Hollnagel makes a fallacious argument that engineers believe in "bimodality," i.e. that things either go right or go wrong (discussed later). This argument, of course, is another strawman as clearly the engineering terminology shown in Figure 2 does not divide behaviors that way. Using engineering terminology, there are "things" that happen (system behaviors) that are not "right" or "wrong" but merely not of much importance in this particular system.

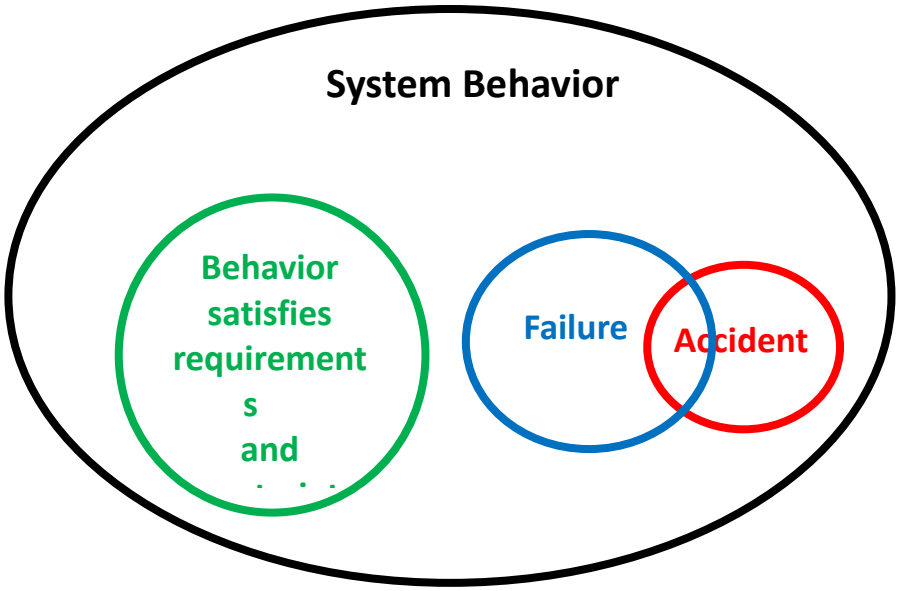


Figure 2. The terminology used in engineering

There is no way to communicate satisfactorily with such sloppy use of language, although it is useful in making fallacious arguments. So let's consider more carefully the terminology engineers actually do use and the problems that arise from using Prof. Hollnagel's terminology.

The first thing that happens in creating any system is that the designers and stakeholders agree on the goals for the system. These may be very general, such as "Produce chemicals" or "Create a rover to send to Mars." Once the goals are agreed upon, they will normally be restated as more detailed "shall" statements or requirements, e.g., "The plant shall produce X pounds of ethylene per day," "The rover shall be able to identify the minerals in the surface of Mars" or "The rover shall be able to travel 5 miles in 24 hours." While the goals may not be achievable because they are often vague and only aspirational, the requirements by definition translate the goals into statements that are detailed enough that it is possible to determine whether the requirements (the "shall" statements) have been achieved or not.

Along with goals and requirements, constraints on how those goals can be achieved are identified. For example, producing the ethylene must not result in explosions, pollution of the environment outside the plant, or the death or serious injury of plant employees. Or the rovers must not pollute the planet surface or send back incorrect information. The same is true for workplace safety: the goal is to produce products or to provide services. The safety constraints are that workers (and customers in the case of services) are not injured in the process. It is possible to increase factory output—achieve the system requirements and "make more things go right"—but at the same time cause repetitive stress injuries in the employees. An example of a basic conflict for healthcare was provided earlier.

Safety is usually associated with constraints rather than goals or requirements. However, that is not always true. If a goal of the system is to ensure safety, the requirements will also involve safety. As an example, one goal of an air traffic control system is to ensure that aircraft maintain a safe distance from other aircraft, obstacles, and dangerous weather conditions. When changing this goal into a measurable requirement, the definition of "safe distance" would need to be specified. Air traffic control systems usually also have requirements unrelated to safety, for example, maximizing throughput in the controlled airspace. Constraints, of course, can also be related to system properties, behavior, or losses that have nothing to do with safety, e.g., must not damage the reputation of the company or must not produce litigation against the company.

Prof. Hollnagel writes:

> Page 177: *The focus of Safety-I is on things that go wrong and the corresponding efforts are to reduce the number of things that go wrong. The focus of Safety-II is on things that go right, and the corresponding efforts are to increase the number of things that go right.*

It is unclear what "things go right" or "things go wrong" are in this context. "Things that go right" may include the goals or requirements but they may also include behavior that the designers and managers of the plant do not care about, for example, the plant contributes to the happiness or contentment of the employees. In other words, the operation of the plant may satisfy goals that were not stated in the requirements for the plant. In the same way, many things can "go wrong" that have nothing to do with safety or accidents, e.g. the working hours interfere with the success of the managers' marriages. It is possible to increase the number of things that go right without ever impacting safety or, in other words, without having any impact whatsoever on the number of things that go wrong. This point is critical. In Prof. Hollnagel's books, there is an assumption that increasing the things that go right will decrease the things that go wrong, i.e., that these are the only two results that can occur and they are a duality. Clearly they are not.

In addition, the focus in safety engineering is not on things that go wrong but on eliminating or reducing hazards. Many things can go wrong without them having any impact on safety or on hazards.

Safety engineering tries to ensure that the safety-related requirements and constraints are satisfied, both of which would normally only be a small subset of "things going right." Engineers, whether concentrating on safety or on other system properties, do not spend much or any effort on all the things that can go right but are not requirements of the system. The set of things that can go right could be practically infinite. In addition, focusing on things going right (the focus of Safety-II) could, and likely might, have no impact at all on safety. This is more than playing with words. The vagueness of the terminology makes some of Prof. Hollnagel's supposedly logical arguments fallacious or untrue. Consider:

> Page 136: *It is more important—or should be more important—that things go right than that things do not go wrong*.

Is it more important that passengers enjoy their flight (a thing that can go right) than that planes do not crash (a thing that can go wrong)? Is it more important that the plane arrives on time than that the passengers arrive safely?

As I said, this is not just playing with words. If we use the correct words then it becomes apparent how incorrect his arguments about Safety-II are. For example, let's say that in the context of safety, what "goes right" is that there is no accident and what goes wrong is that there is an accident. The statement above then, using substitution, becomes:

> "It is more important—or should be more important—that there is not an accident than that there is not an accident."

Or more specifically, it is more important that your car brakes stop the car when necessary to avoid an accident ("go right") than that your brakes do not fail to stop the car when necessary to avoid an accident (do not "go wrong"). By using Hollnagel's vague and undefined terminology, it is possible to create nonsensical statements or truisms. Safety engineering, and engineering in general, would never do this.

In a different type of example, "go right" might be that the plant produces its daily quota of chemicals while "go wrong" might be that toxic chemicals pollute the environment around the plant. That would translate to it is more important that daily quotas of chemicals are produced (go right) than that the plant does not release toxic chemicals into the environment (does not go wrong). They could both happen and the relative importance is a trans-scientific question. Engineers avoid these semantic problems by defining hazards and setting the goal in safety to be eliminating or controlling hazards.

In his second book, where he summarizes the argument in his first book, Prof. Hollnagel says:

> 7(2): *Efforts to try to prevent something from going wrong should therefore be replaced by efforts to ensure as far as possible that everything goes well. This leads to the definition of Safety-II as the situations where as much as possible goes well.*

This is exactly what is done in systems where there are many accidents. The designers and operators concentrate on the non-safety goals, i.e., emphasize achieving goals unrelated to safety—where things "go right"—and put little effort into preventing accidents, that is, they don't put effort into avoiding things "going wrong." It seems incredibly dangerous to me.

The bottom line is that it is not possible to determine if things went "right" with respect to safety without defining what "going right" means with respect to hazards or accidents, i.e., "going wrong."

There is an additional problem. If "things" do "go right"—perhaps nothing untoward happens immediately—how do we know that an accident could not have occurred under slightly different circumstances? Simply focusing on behavior when things go right (even when goes right is defined in terms of hazards) could be dangerously misleading about what is required for safe behavior.

Prof. Hollnagel does raise this question (on page 135) about how we can know or can see what goes right. But his only explanation of how to do this is a quote from a Sherlock Holmes story that seems to have nothing to do with product or system safety. Or even workplace safety.

> "*You can only notice—and know—that something is wrong if you know what should have happened in the everyday case." The understanding of how something is done (everyday work) is a necessary prerequisite for understanding whether something is (potentially wrong)*."

This is exactly why engineers use specifications, including requirements and constraints, to identify what will be defined as "going right" in the system. In addition, it is absolutely <u>not</u> necessary to understand how all work is done in order to identify hazardous behavior. Hazards and hazardous behavior can be defined independent of understanding how everyday work is done.

In fact, an equivalent statement to the one above is to say that if we understand how hazards can occur, then we have, by default, defined how "everyday work" should be done to prevent them. If a hazard can occur when an operator leaves a valve open after doing maintenance, for example, then I know that in order to avoid that hazard the operator must always close the valve after maintenance work or I need to redesign the system so that leaving the valve open is not hazardous. I don't need to understand how all maintenance is performed on the system, including that unrelated to the valve or to that part of the system.

Now that making "things go right" has been shown to have little to do with safety, let's consider the things that can "go wrong." Surely, *that* must have something to do with safety. However, as with the things that can go right, there are an extremely large number of things that can go wrong that we don't care about, i.e., they are unrelated to the system requirements or constraints. Within this large set of things that can go "wrong," there are two subsets that can be called failures and accidents. Because anything that goes wrong could be declared to be a "failure," even if it has nothing to do with the requirements for the system, engineers define failure in terms of the system specification, i.e., a *failure* is the nonperformance or inability of the system or component to perform its <u>specified</u> function for a <u>specified</u> time under <u>specified</u> environmental conditions.

The word "specified" is important. Sometimes, the word "intended" is used, but the problem is "intended by whom"? Who determines what type of behavior is intended? Instead, what is "intended" has to be identified in the requirements and constraints and thus be specified. Without such a specification, we could not know what to build or whether we achieved what was intended by those who commissioned the system. And when "intended" is only defined in retrospect (e.g., after a loss), then it becomes useless as an engineering goal because the engineers could not have possibly designed the system to prevent it without knowing what it was during the design process.

Finally, there is a problem with "as possible," as in Hollnagel's definitions of "as few things as possible go wrong" vs. "as many things as possible go right." Again, this phrase is undefined and undeterminable. How do we know what is possible? Does it mean we've put a lot of effort into eliminating it? How much effort is enough? When and how do we determine that we have as few things "as possible" going wrong or as many things "as possible" going right? How do we determine success?

Finally, let's look at Prof. Hollnagel's conception of causality in Safety-I as shown in his Figure 3.2 (shown here as Figure 3), which he labels the "hypothesis of different causes. There are many problems with this figure (and most of the figures in his books). The first is that Prof. Hollnagel seems to see engineering decisions and the operation of engineered systems as black and white. It almost never is. We can ignore the two components on the right, labeled as "acceptable outcomes" and "unacceptable outcomes." These terms are undefined and therefore not useful. It may be acceptable to the CEO of an airline if the planes are frequently late as long as fuel is saved and profits therefore are increased. That may be completely unacceptable to the passengers. Engineers do not use such undefined terminology.

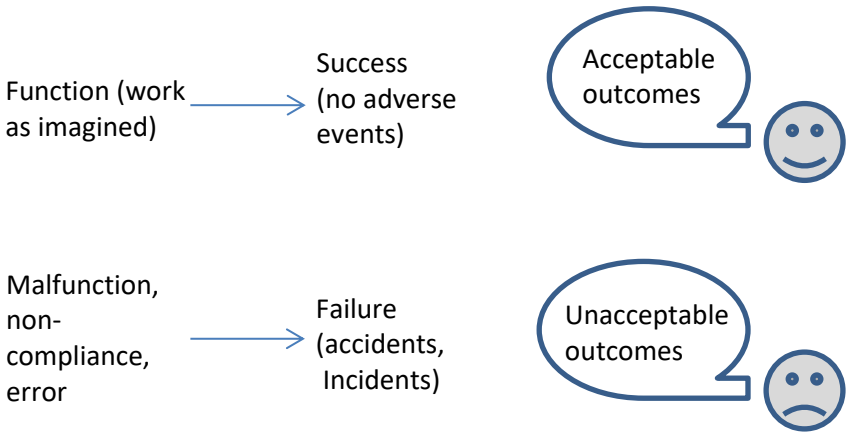


Figure 3 (Hollnagel Figure 3.2 on Page 50): "Hypothesis of different causes"

The other labels are equally as useless. The figure (3.2 in Hollnagel and 3 here) contains one of the most important oversimplifications and misuse of terminology that occurs throughout his books and his figures in order to create his strawman "Safety-I." The figure says that if the system works as "imagined," then there will be no accidents. The problem is that "works as imagined" is again undefined. It is a term used frequently by social scientists, but not by engineers. If "as imagined" means without accidents (or "success" using the meaningless and undefined term used here by Prof. Hollnagel), then the statement is a truism and meaningless. Basically, it says if I don't want accidents, then if I don't have accidents my system is successful. If I do have accidents, then my system is unsuccessful (unacceptable). If by "work-as-imagined" he means not wanting accidents, then the figure says that I won't have accidents unless accidents occur and I have them.

Often, "work-as-imagined" in social science is used to describe the fact that workers do not always follow procedures. Using this definition, however, the figure makes even less sense. The behavior of a system or a worker can deviate from the specifications without there necessarily being an accident. In fact, workers may deviate from procedures to *prevent* an accident. That may be the point that Prof. Hollnagel means to convey in this figure by showing a simplistic strawman argument that purports to be the supposed "philosophy of Safety-I" (Page 50 and throughout his books). Because Safety-I does not exist, it is hard to argue with this strawman diagram and description. He posits Safety-I to be stupid and then draws a stupid diagram to prove it. The only possible rebuttal (and the important one) is to point

out that both his strawman and the diagram have no relation to reality in engineering. His definition of this as "different causes" is discussed later in this paper.

Figure 4 instead shows, using engineering terminology, the true philosophy of causality in safety engineering as it exists today and has for at least a century: The conclusions about how safety engineering conceives of causality become very different than implied by Prof. Hollnagel in his oversimplified and misleading Figure 3.2.

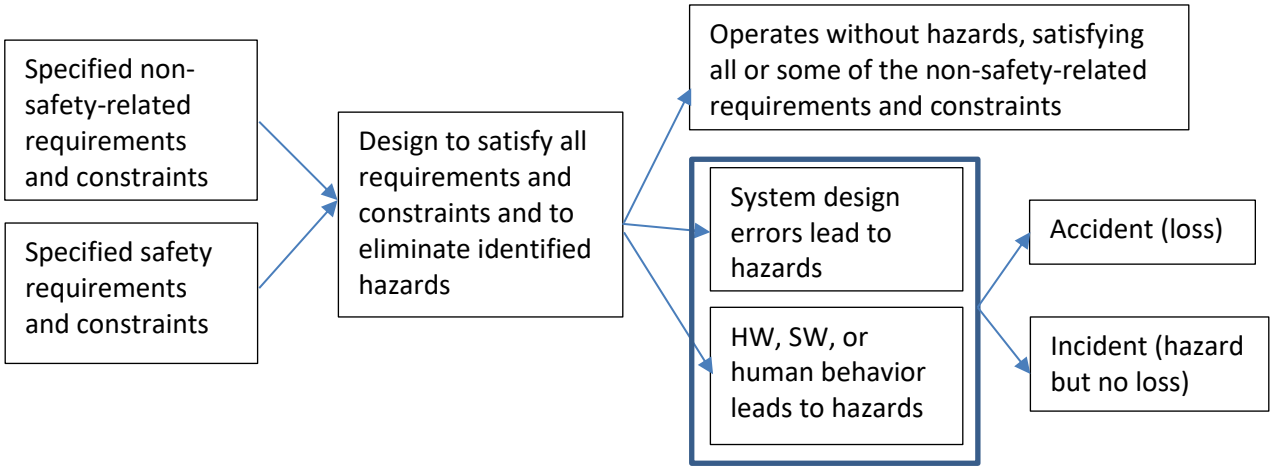


Figure 4: Causality in System Engineering

Engineers start by specifying requirements and constraints, both safety-related and non-safety related. They then design the system—including designing the human-automation interface, human controls, and human procedures—to satisfy all the requirements and constraints and to eliminate or mitigate any identified hazards. One potential result is that the system operates without hazards, while satisfying some or all of the non-safety-related requirements and constraints. There may be "failures" in this case (i.e., not all the non-safety requirements and constraints may be satisfied) but there are no accidents or incidents. Accidents and incidents result from hazards, which in turn result either from system design errors or from component hardware, software, or human behavior (which may or may not involve failures) that leads to hazards. Finally, hazards can lead to accidents or incidents, depending on factors outside the system boundary.

The automotive engineer spends most of his or her time designing the braking system to stop the car under the circumstances specified in the requirements and constraints. Because engineers know, however, that designs and behavior of the designed systems may not always be perfect, they at the same time try to eliminate or mitigate any hazardous behavior that may occur, for example in hydraulic braking systems, by providing a way to stop the vehicle even if hydraulic pressure is lost.

The crux of the problem with Hollnagel's diagram (Figure 3) and his argument is his claim that "Safety-I" hypotheses multiple causes of accidents while "Safety-II" says there is only one cause of all behavior. The book confuses the cause of a particular accident (for example, a leak of brake fluid) with generic models of accident causality that identify how all accidents occur. In fact, both safety engineering today and Safety-II each are based on one generic causality model and thus on one generic cause of accidents. But that discussion needs to wait until after causality models are discussed below.

Notice that Prof. Hollnagel in his diagram (Figure 3) assumes that all failures, non-compliance, and errors lead to accidents or he thinks that others believe this (it is hard to tell which from his frequent use of strawman arguments). This assumption about all failures leading to accidents is simply not true. It is also not true that accidents occur only if there are failures, non-compliance, and errors.

As the diagram in Figure 2 shows, there is an overlap between failures and accidents, but a system that is completely reliable is not necessarily safe and one that is safe does not need to be reliable. Accidents can occur without any component failures. A hazard is not the same as a failure nor necessarily the result of a failure—or even a "latent failure" (as used by Reason, see the Swiss cheese model below), whatever that means.

Understanding the difference between reliability and safety is so important in safety engineering today that it is worth delving into the topic more deeply. Things can go "wrong" and the system can still be safe and the converse is also true, things can go "right" and major losses can occur.

Safety is a Different System Property than Reliability

Confusion between safety and reliability is very common. While I don't know for sure, I speculate that it may come from the fact that until recently reliability was in many cases a reasonable proxy for safety. That is, making systems and products more reliable by making their components more reliable did make them safer because most accidents involved component failures. That is no longer true. In today's world, reliability and safety are different system properties and sometimes even conflicting.

System components can operate 100% reliably and accidents may still result, usually from unsafe interactions among the system components caused by system design errors or simply by complexity that overwhelms our ability to identify and thus handle all potential unsafe component interactions. Examples are shown throughout this paper.

In addition, the larger environment beyond the system boundaries (i.e., context) is important in determining whether something is safe or not. As a simple, real-world example, consider going out to the middle of a large deserted desert, pointing a gun away from oneself, and firing. If there is nobody or nothing in the vicinity, the gun and the activity could be considered to be both reliable and safe. Consider, however, doing the same thing in a crowded mall. The gun has not changed, the gun's reliability has not changed, and the action (pulling the trigger) has not changed. But the safety certainly has.

The safety of something can only be determined by considering the context in which the thing operates. This is true for components as well as systems. In the Ariane 5 spacecraft loss, which was directly related to the Inertial Reference System (IRS), the same IRS operated safely in the Ariane 4. When it was reused in the Ariane 5, it contributed to the explosion and loss on its first launch. The IRS did not change. What changed was the design of the system in which the IRS was used. In particular, the Ariane 5 had a higher launch trajectory than the Ariane 4.

Lots of accidents today occur in systems where there are no component failures—for examples, see Chapter 1 of the STPA Handbook [Leveson and Thomas, 2018]. An example used throughout this paper occurred when an A320 aircraft was landing at Warsaw airport in a bad storm with heavy rain. The aircraft did not decelerate in time and crashed into a small mound at the end of the runway. The flight crew tried to activate the reverse thrusters (i.e., the temporary diversion of an aircraft engine's thrust so that it acts against the forward travel of the aircraft) which is the normal way to decelerate, but the software would not let them do so because the software "thought" that the aircraft was still in the air. It is dangerous to activate reverse thrust when airborne, so protection had been built into the software to prevent this from happening, i.e., to protect against a flight crew error. But the engineers did not fully account for all possible environmental conditions at an airport, including the rare ones. In this case, the conditions indicating the aircraft had landed did not hold for unusual reasons.

Design errors of this type are theoretically identifiable during system development. When the systems we were developing were simpler (mostly before the computer age in engineering), the designs could be exhaustively tested and engineers could usually think through all the potential interactions during design. These two properties are no longer true for today's systems. We are building systems for

which we cannot anticipate or guard against unintended behavior. Complex systems today have so many potential states that exhaustively testing them before use is simply not feasible. The result is an increasing number of what engineers call the "unknown unknowns," i.e., unanticipated things that happen during use of a system. We can no longer assume that design errors will be detected and removed before use of a product or system.

Considering reliability only at the system level, instead of the component level, does not help. Complex systems almost always have many requirements (or goals) along with constraints on how those goals can be achieved. As an example, a chemical plant may very reliably produce chemicals (the goal or mission of the plant) while at the same time polluting the environment around the plant. The plant may be highly reliable in producing chemicals but not safe.

In summary, safety is a system (emergent) property—that means safety is only determinable by examining the behavior of all the components working together and the environment in which the components are operating. System safety can almost never be determined by simply looking at the behavior of an individual system component, including the human operators.

Another way of saying this is to go back to where this section started, i.e., that a system component failure is not equivalent to a hazard. Component failures can lead to system hazards, but a component failure is not necessary for a hazard to occur. In addition, even if a component failure occurs, it may not be able to contribute to a system hazard. Therefore, focusing on either increasing the things that "go right" (increasing component non-failures, which is equivalent to decreasing component failures) or even decreasing the things that "go wrong" (decreasing component failures) may lead to no impact on safety whatsoever.

To summarize this section so far, Safety-I does not exist, i.e., it is a misrepresentation of engineering in general and safety engineering specifically. Safety-II, on the other hand, violates what we know about engineering and is more likely to decrease safety or, at best to have no impact, than to increase safety. Safety-III is, I believe, the best path forward. It defines safety as freedom from unacceptable losses as identified by the system stakeholders. The goal is to eliminate or control hazards, which are the states that can lead to these losses.

To examine other aspects of what Prof. Hollnagel defines as Safety-I and Safety-II, it is necessary to clarify some more definitions. Most of Prof. Hollnagel's arguments for Safety-II rest on nonstandard definitions of standard terminology. His two books, in fact, are filled with undefined terms or uniquely defined ones.

Most of his arguments seem to stem from his basic assumption that

> *The thinking that was relevant for the work environments of the beginning of the twentieth century is unlikely to be relevant today when socio-technical systems are not decomposable, bimodal, or predictable.*

Determining whether this statement is true will depend on the definition of the terms in it, i.e., system, sociotechnical system, composition, bimodality, and predictability.

What is a System?

Let's start with the basic definition of a system. It underlies all understanding of system safety (and system engineering). Prof. Hollnagel defines it as:

> *The traditional definition of a system has therefore been with reference to its structure, i.e., in terms of the parts and how they are connected or put together. A simple version is to declare that a system is anything that consists of parts connected together. A slightly more elaborate one is that a system is a 'set of objects together with relationships between the object and between their attributes.' [p. 97]*

While this may be a common informal definition of system, it is not the definition used in engineering or in systems theory, which is:

> *System*: A set of things (referred to as system components) that act together as a whole to achieve some common goal, objective, or end.

The goal is the critical part of the definition. Prof. Hollnagel is not the only one who leaves out the "goal" or "objective" and states essentially that a system is a connected set of things or parts forming a unified whole. Note that the connections are not mentioned at all in the engineering definition provided above.

My ear, my shoe, and my hair are interconnected through their connections to my body. They could be considered to be a system if a purpose could be conceived for considering these individual things together; a purpose is basic to the concept. As an example that does fit the definition, the human digestive system consists of the gastrointestinal tract plus the accessory organs and hormones used in digestion. The purpose is the breakdown of food into smaller and smaller components, until they can be absorbed and assimilated into the body. The goal of a transportation system is to move people from one place to another.

Some definitions of a system do require, in addition to having a goal or purpose, that the components of the system must be interdependent or connected or interacting, but none of these conditions are really necessary to have a system and they constrain the definition in a way that excludes things that usually are considered to be systems. The system components may be either directly or indirectly connected to each other, with the latter including connections involving the system purpose only and various types of non-linear interdependencies. This property is important when causality models are considered later.

The consequence of this definition is that a goal or objective for the system is fundamental. But there may be different objectives for those defining a system than for those viewing that system. Consider an airport. To a traveler, the purpose of an airport may be to provide air transportation to other locales. To local or state government, an airport may be a means to increase government revenue and economic activity in the area of the airport. To the airlines, the purpose of an airport may be to take on and discharge passengers and cargo. To the businesses at the airport, the purpose is to attract customers to whom they can sell products and services. When talking about a system, it is always necessary to specify the purpose of the system that is being considered.

The critical part of the use of the term "system" is that while the components of a system may exist, in fact

> *A system is an abstraction, that is, a model conceived by the viewer.*

Systems are not labeled as such in nature, they are a construction of the human mind. The physical components exist, but humans impose a structure and purpose on those components allowing them to be considered together as a system.

The observer may see a different system purpose than the designer or focus on different relevant properties. Specifications that include the purpose of the system are therefore critical in system engineering. They ensure consistency of mental models among those designing, using, or viewing a system, and they enhance communication. Notice that different components of the airport may be included in particular "airport" systems, such as an airline view of an airport as passenger check-in counters, ramps to the planes, and taxiways versus a commercial view containing shops and customers.

In summary, "systems" are *models* or *abstractions* laid upon the actual physical world by human minds, not by nature. The components that are considered in any "airport system" or subsystem and the role they play in the system as a whole may be different for each concept (model) of an airport system or of airport subsystems. The basic idea here is that the purpose or goals of the system being considered must be specified and agreed upon by those modeling and analyzing a particular system and that these aspects of systems are abstractions or models *imposed by the viewer* on the real world objects.

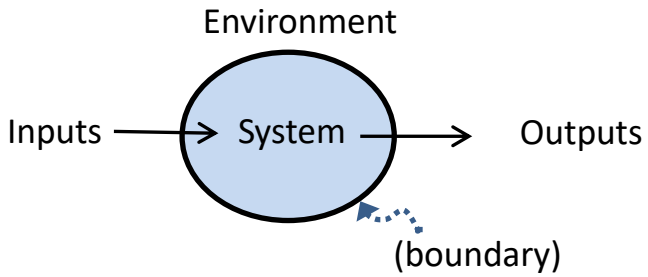
There are some basic assumptions underlying the concept of a system:

- The system goals can be defined and
- Systems are atomistic, that is, they can be separated into components with interactive behavior or relationships between the components, whether the interactions are direct or indirect. (Prof. Hollnagel's claim that sociotechnical systems are not decomposable, i.e., not atomistic, is discussed after "sociotechnical" systems are defined below.)

Systems have *states*. Informally, a state is a set of relevant properties describing the system at any time. Some properties of the state of an airport viewed as an air transportation system may be the number of passengers at a particular gate, where the aircraft are located and what they are doing (loading passengers, taxiing, taking off, or landing). The components of the state that are relevant depend on how the boundaries are drawn between the system and its environment.

In safety, some subset of the total number of states of the system may be defined as "hazardous." Examples of hazards for an airport is the state where an aircraft is not able to decelerate when headed toward a ramp, food is served in airport restaurants that is injurious to human health, or even that large numbers of passengers are unable to reach their departing aircraft because of delays in security screening. What are considered as hazardous states depends on the "losses" defined by the stakeholders.

The *environment* is usually defined as the set of components (and their properties) that are not part of the system but whose behavior can affect the system state. Therefore, the system has a state at a particular time and the environment has a state. The concept of an environment implies that there is a boundary between the system and its environment. Again, this concept is an abstraction created by the viewer of the system and need not be a physical boundary. What is part of the system or part of the environment will depend on the particular system and its use at the time, i.e, the definition of a system boundary may be useful for the current purpose but may be changed when considering a different purpose even though the system is comprised of some of the same components, as in the airport example.



System *inputs* and *outputs* cross the system boundary. This model or type of system is usually called an *open* system. There is also a concept of a *closed system* (which has no inputs or outputs), but this concept does not have much relevance for the engineered systems with which we are most concerned in system safety.

Now that the basic definition of a system is clarified, let's consider further Prof. Hollnagel's statement at the beginning of the section:

> *The thinking that was relevant for the work environments of the beginning of the twentieth century is unlikely to be relevant today when socio-technical systems are not decomposable, bimodal, or predictable*

While it is not surprising that the approach to work environments (which I assume refers to workplace safety) from 100 years ago is not relevant today, the reasons given in the above statement make no sense to me. We need to look at Prof. Hollnagel's definition of *decomposable*, *bimodal*, and *predictable*

first along with the standard definitions for these terms in engineering and systems theory along with the definition of *sociotechnical systems* to understand why and to understand what Safety-III is.

Sociotechnical Systems

All useful systems are sociotechnical. Ralph Miles Jr., an early practitioner of systems theory, wrote:

> *Underlying every technology is at least one basic science, although the technology may be well developed long before the science emerges. Overlying every technical or civil system is a social system that provides purpose, goals, and decision criteria [137, p. 1]*

No technical systems exist in a non-societal vacuum. However, a problem that is rampant both among engineers and social scientists is that they focus on either the social aspects or the technical aspects and not their interaction, that is, they define what is inside the system very narrowly. In other words, engineers often draw the boundary of their system of interest around the technical components of the system and leave the social components in the environment. They then focus on the parts of the system within the boundaries they have drawn. It's not that they don't know the social elements are important, but they leave those to social scientists.

Social scientists do use the term "sociotechnical systems" more than engineers, but then they almost always focus all their attention on the humans and social structures, as does Prof. Hollnagel, and don't examine the tight connections between the two. That is, they draw their system boundaries around the social or human aspects and put the technical aspects in the environment.

As just one example, Prof. Hollnagel states:

> Page 111: "*We develop systems we are unable to control—sometimes even deliberately so, as in the case of dynamically unstable fighter aircraft*."

Who is "we" here? The first "we" seems to be "society" or "engineers" but the second "we" must mean only human controllers (pilots). In fact, dynamically unstable aircraft are controlled, but they are controlled by computers. Otherwise, they would crash every time we tried to fly them. Computers and other means of control (beyond human controllers) are omitted in Prof. Hollnagel's arguments for Safety-II. Safety-II is not a sociotechnical approach because the technical is almost completely ignored.

The problem is that when focus is primarily only on the human operators and "human error," or even on only the technical components, then we have too small a keyhole to allow us to understand and solve our problems, including improving safety. The impact of the technical design and the social systems on each other, i.e., the integrated sociotechnical system behavior, must be considered to deal with safety in our complex world today.

For example, the design of cars is not simply related to the physical components of the car but also related to the design of the road system along with other factors such as regulations. Regulations, such as fuel mileage rules, relate to social issues and not only to the technology involved in designing cars. At the same time, our social structures are related to the limits and requirements of our technology. Human error when interacting with a technological system is almost always a function of the design of the system within which the human is working. Human factors experts often focus on the impact of the system design on human error. A good introduction to this topic can be found in Don Norman's books such as *The Design of Everyday Things*. Prof. Hollnagel, in his two books on Safety-II used here, focuses on the impact of human error on the system design and ignores the other side of the coin, namely, *the impact of system design on human error and behavior.*

To reduce human error, we need to design our systems with respect to both. We cannot just focus on the human operators. In the same way, to improve human adaptability within systems and system resilience, we need to design the context in which the human works to allow positive adaptability—humans can adapt in the wrong ways, and, in fact, often do. Jens Rasmussen has hypothesized that

systems migrate to states of higher risk under competitive and other pressures [Rasmussen, 1997]. To improve system resilience, we need to design the system *as a whole* to be resilient, not just focus on one part, such as the human operators.

In fact, the role of human operators in systems is at best staying steady or decreasing, the role of hardware is mostly decreasing, the role of designers and managers is for the most part staying steady, and the role of software is increasing. It makes no sense to concentrate only on the role of humans (actually Prof. Hollnagel seems to focus on "human operators," not all humans in systems) in trying to make systems safer unless we believe that all accidents are the fault of human operators.

Any view of safety has to include all the aspects that impinge in any way on safety, which for today's highly coupled systems, is usually all of them. Thus, improving safety requires focusing on the sociotechnical system as a whole and designing it to achieve our goals, including the defined safety goals. One of the reasons "unintended consequences" so often happen is that we focus on optimizing one part of a complex, sociotechnical system without considering how it will impact the behavior of the other parts.

While it is not practical to train engineers as social scientists or social scientists as engineers, we can develop tools that allow them to work together on complex sociotechnical system problems. Safety-III emphasizes this by providing engineering modeling and analysis tools that include and operate on the entire sociotechnical system as a whole.

Prof. Hollnagel does acknowledge the interdependence between social and technical factors

> Page 183*: In a socio-technical system, the conditions for successful organisational performance – and, conversely, also for unsuccessful performance – are created by the interaction between social and technical factors rather than by either factor alone.*

But the definition of Safety-II ignores the technical factors and focuses on human operator behavior alone. There is almost no mention in these two books of the design of technology. Humans operate within systems and their actions are limited by the systems in which they work. A true sociotechnical or "systems approach" views all the components of the system, human and otherwise, as a whole. Safety problems cannot be tackled successfully by focusing only or primarily on the human operators, as in Safety-II.

Let's go back to the quote above:

> *The thinking that was relevant for the work environments of the beginning of the twentieth century is unlikely to be relevant today when socio-technical systems are not decomposable, bimodal, or predictable.*

Now that we have definitions for system and sociotechnical system, let's look at the other terms he uses (decomposable, bimodal, and predictable) and compare his definitions with the standard engineering definitions.

Decomposition and Emergence

Prof. Hollnagel claims that "*socio-technical systems are not decomposable."* Virtually all systems, whether sociotechnical or not, are decomposable. Without any examples, it is hard to understand what Prof. Hollnagel could possibly mean by this. The basic definition of a system in System Theory, presented above, is that systems are atomistic, that is, they can be separated into components with interactive behavior or relationships between the components, whether the interactions are direct or indirect.

Perhaps his confusion arises from the concept of analytical decomposition. One of the great advances in modern science, starting around the 1600's was this concept, whereby complex things are broken down into their constituent components. The idea was that the individual components could be analyzed separately and then the analysis results combined to get a result for the whole.

Engineered systems are always decomposable; they are created by putting components together. Social systems are also decomposable, particularly the parts that have been designed like management and government. I cannot think of any social system that cannot be decomposed. At the least, they are decomposable into the humans involved and usually the social structures in which the humans operate.

As Prof. Hollnagel repeats this assertion—*socio-technical systems are not decomposable*—throughout the book, at least one example of a non-decomposable system would have been helpful to support this claim. There are some particular aspects of social systems that cannot be decomposed, but not the systems themselves. There is a difference between systems and the properties that a system may have. For example, safety is a system property, not a system.

The idea of decomposability and analytic decomposition forms the basis of modern physics and engineering, along with many other fields. It was only in the middle of the last century that the complexity of some of the systems being studied or created was so great that many of the previous analytical methods that were created to operate on the decomposed components became less effective. Systems Theory[9] was created to handle these systems, originally in biology (von Bertalanffy) and in engineering and math (Norbert Weiner and others, who called it cybernetics).

In Systems Theory, it is not that systems cannot be decomposed, but that there are some system *properties* (not systems or components) for which the composition (combination) of the analysis of the components cannot be combined for a "system value." These are called emergent properties. They arise from the interactions among the components and not simply from the combination of the values for individual components.

An example of an analytically decomposable property for an aircraft is weight. If we add up the weight of all the components in the aircraft, we get an acceptable value for the total weight of the aircraft. We can even get good estimates of a non-physical property by combining the measures of those properties for each component into a combined measure for the system as a whole. As an example, for simpler systems (but not very complex ones) the reliability of the individual components and information about how they interact in the system can be used as a way to calculate the reliability of the system as a whole.

Clearly, as stated, engineered systems are decomposable. We create them by putting components together. Social systems, as stated, are also decomposable. Remember, a system is an abstraction formed in the mind of the observer of that system. The airport example purposely was provided because it is easy to view it as being a sociotechnical system. In that example, the airport was decomposed into the subsystems for commercial businesses and air transportation. We might decompose the businesses into restaurants, merchandisers (books, clothes), and services (e.g., massage parlors). We could decompose the air transportation subsystem into air traffic control, aircraft, checkin and ramps, security gates, etc. A different decomposition is to decompose the airport into the airlines that operate there. Or the airport could be decomposed into money-making entities versus other services (e.g., security). Remember, a system is an abstraction created in someone's mind; the components exist in reality but the concept of a system is something that we impose on those components.

If we look at social systems, we may decide to focus on different parts of them, such as the people involved, the social structures (government agencies and political entities), or other aspects. We can decompose these and describe properties of the individual components (such as properties of the individual humans involved) or system properties that describe properties created by the interactions among the components such as democratic or authoritarian.

[9] Note that Systems Theory and Complexity Theory are not the same although sometimes people confuse them. The difference is described briefly in Appendix A.

The theoretical foundation of analytic decomposition is based on assumptions about the system that allow the synthesis or combination of individual component analyses to analyze system properties. The primary assumption is that the separation of the whole into separate components does not distort the phenomenon of interest. For this assumption to be true, the following sub-assumptions must be true:

- Each component or subsystem operates independently.
- Components act the same when examined singly as when playing their part in the whole
- Components/events are not subject to feedback loops and nonlinear interactions.
- Interactions can be examined pairwise.

These assumptions are not usually true for complex systems today.

As an example of a system property that cannot be analyzed using analytical decomposition, consider gridlock on the roads. Gridlock cannot be predicted by simply calculating some properties of individual automobiles and then combining the results. Whether gridlock occurs will depend on such system factors as the number of lanes on the highway, the time of day, the number of people who want to traverse the highway at various times of day, whether previous experiences with gridlock convince enough people to travel at non-peak times or take alternative routes or transit, whether there is a global pandemic occurring and stay-at-home orders have been issued, etc.

The appropriate conclusion is NOT that gridlock or other system properties cannot be predicted and analyzed, as Prof. Hollnagel assumes in these two books. It is simply that doing so requires more than analyzing the individual components (e.g., a car) and combining them in a straightforward manner by looking only at their physical or logical interfaces with other components (the other cars). Analyzing gridlock and predicting it requires more sophisticated tools. This need is exactly why Systems Theory was created, as explained later.

For complex systems, some or all of the above assumptions required for analytic decomposition to be effective may not hold for some properties. The properties that do not satisfy the assumptions are called system or *emergent* properties. Safety is a system (*emergent*) property. If I examine a system component in isolation from the others, there may be some hazards associated with it, for example, sharp edges or toxic chemicals or materials. But the system operating as a whole may have other emergent hazards that arise in the interactions among the components.

For example in the airport example, the flight of a particular airline being late does not depend simply on factors related to that airline but on complex interactions among many other systems at the airport and around the country, e.g., air traffic control and congestion at the airport to which the aircraft is going or came from, crew duty cycles, weather, etc. And it does not depend only on a simple combination of these, but also on how they interact with each other. The hazard when driving of crashing into the car in front depends not only on the brake system components and how they interact (which given today's use of digital components and the interactions between the braking system and other auto systems is complex by itself), but also on driver reaction time, weather and visibility, driver distraction, the behavior of other drivers on the road, etc.

Newer tools for dealing with safety using Systems Theory provides us with the ability to deal with this level of complexity, but simply looking at the flexibility and adaptability of the individual drivers (Safety-II) and how they were able to avoid an rear-end collision in the past or how they operate their individual cars will not provide the answers we need. Safety-III is founded on Systems Theory and thus can deal with emergent properties like safety in the entire sociotechnical system.

Understanding emergence is critical to understanding and dealing with safety in complex, sociotechnical systems. Prof. Hollnagel writes:

> Page 128: *There are, however, a growing number of cases … where it is impossible to explain what happens by means of known processes or developments.*

In 40 years, I have never found an accident where it was not possible to explain what happened by means of known processes or developments beyond a very small number of accidents that were never investigated properly (although they were usually able to blame the operator anyway) or those in the past that involved some poorly understood physical principles at the time. An example of the latter is the steam engine explosions in the 19th century before the science of thermodynamics was understood.

Not completely understanding the underlying science does not mean that there is no way to prevent these accidents. Indeed, as an example, many devices were created to prevent accidents in steam engines such as relief valves and other design and operational controls. Once the science of thermodynamics was better understand, the prevention of steam engine explosions became more effective, but we still were able to engineer safety devices to prevent or reduce the impact of steam boiler explosions before completely understanding the underlying science.

Similar examples exist today when people push to use a technology before it is fully understood. It is probably less common today, at least in safety-critical systems where we have social and legal controls (including liability) that tend to limit such use. I wish Prof. Hollnagel had provided an example of a case where it is impossible to explain what happens by means of known processes or developments.[10] He makes this claim many times in his book but without any examples. We might not be able to fully understand why something happens—although even that is rare in critical engineered systems—but we almost always understand what happened and usually can predict and prevent it using new sophisticated tools such as Systems Theory and Safety-III techniques.

Page 132: *Since emergence cannot be explained in terms of causality …*

Emergence in Systems Theory does not imply that there is no explanation possible in terms of causality, only that the explanation is different than for the events or properties that are not emergent, i.e., the explanation is in terms of the interactions among system components rather than the behavior of individual components.

Page 131: *In the case of emergent outcomes the causes are, however, elusive rather than real.*

Causes can be real and be elusive and vice versa. These are not opposites.

He later changes his argument about emergence:

> Page 128: It *is still possible to provide an explanation of what happened, but it is a different kind of explanation. In such cases the outcome is said to be 'emergent' ('rising out of' or 'coming into existence) rather than resultant ('being a consequence of'*). *The meaning of 'emergence' is not that something happens 'magically,' but simply that it happens in a way that it cannot be explained using the principles of linear causality.*

Contrary to Prof. Hollnagel's statement, emergent properties <u>are</u> resultant in that they result from the interactions of the system components. That is, in fact, the definition of emergence. They can usually easily be explained in terms of why they happened. For example, I can explain why gridlock occurs on Highway I-5 in Seattle every weekday at rush hour and why accidents occur in complex sociotechnical systems.

Prof. Hollnagel is correct that they cannot be explained using linear causality, but linear causality, while the basis for most existing causality models, including his own, is not the only type of causality explanation that does not involve "magic." The topic of causality, and in particular linear causality, is so important in safety that it is covered in depth below.

Prof. Hollnagel continues:

---

[10] Even in the mysterious disappearance of Malaysian Airlines Flight 370, we have several plausible explanations in terms of known processes and developments for what actually happened. The problem is that, without ever finding the wreckage, it is difficult to select between these possible explanations.

"Indeed, it implies that an explanation in terms of causality is inappropriate---and perhaps even impossible.

I'm not sure where Prof. Hollnagel got the idea that emergent properties don't have causes or that their causes cannot be described and understood. The Warsaw reverse thruster accident can be easily explained, as it was above. It just cannot be explained by simply combining properties of each individual component but requires examining their interactions (which is the definition of emergence). Prof. Hollnagel writes:

This is typically the case for systems that in part or in whole are intractable. The first known use of the term was by the English philosopher George Henry Lewes (1817-1878), who described emergent effects as not being additive and neither predictable from knowledge of their components nor decomposable into those components. Pg. 129

Prof. Hollnagel seems to be interpreting what Lewes said here differently than I do. What Lewes said appears to agree completely with the standard definition of emergence as I have described it. In fact, he uses the term as related to "effect," which is equivalent to what I have called properties. It is neither inappropriate nor impossible to explain the cause of accidents arising from emergence, i.e., from interactions among the components rather than failures or behaviors of the individual components. Several examples have been given in this paper. It sounds like Lewes was talking about system properties rather than explaining accident causality. As I said, the fact that a property is not additive or that it requires an examination of the interactions and not simply the combination of individual components, does not make an explanation in terms of causality inappropriate or impossible.

There is a very strange statement about emergence on page 132 of the first book:

Page 132: *In the case of emergent outcomes, the 'causes' represent patterns that existed at one point in time but which did not leave any kind of permanent trace.*

While clues about causes of accidents may be difficult to find, for example, when they are related to software, I cannot imagine what this might have to do with emergence, even using his definition of emergence.

"Bimodality"

Once again, let's look at Prof. Hollnagel's basic thesis:

The thinking that was relevant for the work environments of the beginning of the twentieth century is unlikely to be relevant today when socio-technical systems are not decomposable, bimodal, or predictable.

The term "bimodality" is used in statistics to describe a continuous probability distribution with two different modes, which appear as two distinct peaks in the probability density function. The standard definition makes no sense in the context in which Prof. Hollnagel uses it. It appears from the description provided that he may be referring to "duality" instead, so let's assume that is what he means, i.e., that there are only two parts to something, usually opposing. Redefining standard technical terminology makes these two books more difficult to understand than necessary.

On Page 99, Prof. Hollnagel claims that, in Safety-I, components that are part of the explanations for adverse outcomes *"either have functioned correctly or have failed."* Of course, there are no references for this statement nor citations nor examples. It is a classic strawman argument.

There is no definition of "functioned correctly" but let's assume he means they satisfied their specification and "failure" means the opposite, which would be the engineering definition of failure. The problem is that there are many accidents where the components satisfy their specification (and thus have not failed) yet a loss occurs. Component failure is one cause of accidents, but there are many

others, including that the specified behavior is unsafe. Reliability and safety are two different properties as explained above.

Elsewhere he argues:

> Page 127*: The ontology of Safety-II is thus that human performance individually or collectively is always variable. This means that it is neither possible nor meaningful to characterize components in terms of whether they have worked or they have failed, or whether the functioning is correct or incorrect.*

At least here he is making it clear that Safety-II concerns human behavior only, although elsewhere he seems to be implying non-human behavior and the entire sociotechnical system. First, I cannot imagine anyone who believes that human behavior is not variable. This statement seems to harken back to Prof. Hollnagel's false assumption that engineers think of humans as machines. However, human behavior variability does not imply that it is not possible to characterize the human's behavior as "correct" or "incorrect" or, more relevant, safe or unsafe.

It is also strange to talk about humans having "worked" or "failed." What could that mean? The use of the word "components" in the second sentence is also odd. Is he implying that all components in a system are humans? The second sentence does not seem to have any relationship with the first.

He then states that

> *This principle of bimodality represents the commonly held assumption that things function until they fail. Whenever an individual component fails, such as a light bulb that burns out, the component will be discarded and replaced by a new (and usually identical component). The same reasoning applies to composite systems although failures sometimes may be intermittent, especially if complicated logic (software) plays a part.*

Maybe he is *not* just talking about human behavior as a light bulb is certainly not a human. But even so, it is hard to know where to start with this statement that seems to contradict basic engineering principles and practices, but for which there are no references or examples beyond the trivial example of a light bulb. None of these statements in the paragraph above apply to complex, engineered systems nor even to social systems. Prof. Hollnagel seems to imply that this assumption is held by some group, without that group being named nor any citations to give a clue as to who might hold this "common" assumption. Once again, this is a classic strawman argument.

It is also unclear what a "composite system" is in the above statement because nearly everything has parts (as discussed in the section on decomposition and the definition of a system above), even a lightbulb (e.g., filament, wires, inert gas, insulation, etc.). Physical component failures can be intermittent in any system. Talking about software "failures" in the same way as lightbulb failures does not match reality. Software does not fail intermittently. It does not wear out according to some probability distribution and then have to be replaced because it will no longer operate. The major problem here is the one discussed above about the distinction between reliability and safety. Software (and hardware) can satisfy its specification and still be unsafe and vice versa. Actually, most things do not "function" until they fail as stated in the above.

> *But even for composite – or non-trivial – systems, it is assumed that performance basically is bimodal: either the system works correctly (as designed) or it does not.*

Again, who assumes this? Where are some references or examples? Most engineers do not. First, systems and components of systems always work as designed. That may not be what was wanted or specified, but there is no way something can operate in a way that violates its design. Even failures are the result of designs that did not eliminate the potential for that failure. The important question is whether the design is safe or not, not whether it "works correctly" or not. It is ironic that Prof.

Hollnagel's own conception of system behavior is "bimodal," i.e., it "goes right" or it "goes wrong." The engineering definition is not bimodal.

Working "correctly" in the above statement, I assume, means that it satisfies its requirements and constraints. That is what "correctly" means in engineering. As discussed above, there are lots of behaviors that the system stakeholders do not care about and are not included in the system requirements and constraints. In addition, usually at least part of the requirements and/or constraints are satisfied so bimodality doesn't seem to mean much as used here. Does it mean that all the requirements and constraints are satisfied or none of them are? What if some of them are? What if they are partially satisfied? Prof. Hollnagel's statement is an oversimplified and inaccurate view of system engineering and system safety. Engineers do not think this way.

Prof. Hollnagel suggests that as a consequence of "bimodality"

> Page 99*: Notion that there must be a kind of congruence or proportionality between causes and effects. If the effects are minor or trivial—a slight incident or a near miss—we tacitly assume that the cause is also minor. Conversely, if the effects are major—such as a serious accident or disaster—then we expect that the causes somehow match that, or at least that they are not trivial.*

This statement seems to imply that people (whoever "we" is) believe causes have some importance independent from the effects. Again, there are no examples or citations. Without any examples, as usual, it is difficult for me to determine why anyone would believe this notion to be true. Actions (causes) only can be classified in importance with respect to the effects. If a pilot pushes a particular button in the cockpit, say button A when it should not be pushed, is that major or trivial? If it results in a serious hazard, then it is major (using Prof. Hollnagel's terminology). But if there is no or only a trivial impact, then the exact same action is trivial.

Complicating things is that the impact of one action, say pushing button A can (and often will) depend on whether other conditions exist at the same time. Those conditions may be beyond our system boundaries and thus beyond our control. There is no way to determine the importance of the pilot activating button A (a potential cause) without knowing (that is, independent of) the consequences. In fact, it is not even clear that labeling causes as trivial or major makes any sense at all. Who are the "we" who believe such a thing? It is certainly not the way that safety engineers look at system safety. This is another strawman argument.

Prof. Hollnagel continues:

> *(A consequence of that is the assumption that the more serious an event is, the more we can learn from it; this is discussed further in Chapter 8). In both cases, the assumption about proportionality may clearly bias the search for causes, so that we look either for something trivial or something significant.*

Again, without any references, it is difficult to know who Prof. Hollnagel thinks makes this assumption or who "we" is. More serious outcomes are investigated with more rigor only because the investigators do not have the resources to investigate everything rigorously nor would that be a good use of necessarily limited investigation resources. I cannot imagine why the seriousness of the consequences would change the search for causes or result in investigators looking for different causes (except for political or liability issues that are beyond the topic of this discussion). Once again, an event or condition is only trivial or significant with respect to the consequences.

> Page 127*: Safety-II … Human performance individually or collectively is always variable. This means that it is neither possible nor meaningful to characterize components in terms of whether they have worked or they have failed, or whether the functioning is correct or incorrect.*

First, note that here and throughout his books, Prof. Hollnagel assumes that all system components are human. Focusing only on humans is not a systemic or sociotechnical approach and is not going to be

effective in preventing accidents in today's potentially dangerous sociotechnical systems. It seems to indicate human error is the only cause of accidents.

It is always possible in engineering to determine whether components have failed and whether their functioning is correct or incorrect because correctness and failure are defined in terms of the specification of the requirements for the components. And those, in turn, are defined in terms of the system requirements and constraints. The real problem here, however, is Prof. Hollnagel's confusion between reliability and safety, as discussion in depth above. In this whole discussion, and in fact in both of his books, he appears to believe that accidents are only caused by failures.

The bimodality principle of Safety-I is therefore obsolete.

The bimodality principle doesn't exist and never did in engineering, just as Safety-I does not exist. So this non-existent principle did not become obsolete and does not need to be replaced in a current practice (Safety-I) that does not exist. His strawman arguments are simply that. Even worse, Prof. Hollnagel wants to replace it with something that already exists and somehow suggests that replacement will improve safety. It's difficult to understand why it would make an improvement if it is already what most people already do as explained later.

Page 141: *Replaces the bimodality principle by the principle of approximate adjustments.*

Without examples or more careful definitions, it is difficult to understand this statement. The two, even if they existed, seem totally different. The opposite of the bimodality principle as described by Prof. Hollnagel would be that things are partially right or partially wrong. Approximate adjustments (which are never carefully defined in the two books) are a way to achieve a goal, they are not a goal state themselves. In other words, one could have both the bimodality principle and use approximate adjustments to describe how the system state becomes "right" or "wrong." Again, the lack of any examples and the use of poorly defined and ambiguous terminology makes it difficult to figure out if there is any truth underlying his statements.

In addition, "approximate adjustments" sounds like simple feedback control loops, which are standard in engineering. Any controller, whether implemented by humans, hardware, or software, uses feedback control. Basically, feedback control means that a goal or setpoint is selected and the controller adjusts the controlled process to achieve the goal using feedback to determine the current state of the controlled process. The current state of the process is then compared to the goal state and adjustments are made accordingly. What is the difference between this basic engineering adjustment (control) process and "approximate adjustments"? Feedback control is the process underlying most engineering, Safety-III, and Systems Theory. Again, this is examined more carefully after the whole issue of causality and models of accident causality are explained later. But two more terms used in Prof. Hollnagel's arguments for Safety-II needs to be defined first.

Predictability

Prof. Hollnagel writes:

> Page 102: *Assumption: Predictability. The order or sequence of events is predetermined and fixed. …processes that were explicitly designed to produce a given outcome---whether the processes were technical ones such as a power plant, human activities such as in accomplishing a task in a factory or production line; or providing a service.*

This is another piece of his overall strawman argument. Of course, no examples or references are provided. While there are some types of hardware where the design does assume that the sequence of events is fixed (although not necessarily when failures occur), this assumption is clearly not true for software or humans. In fact, software sometimes is purposely written to be nondeterministic. And I cannot believe that any engineer would assume this statement is true about the humans in the system.

Perhaps this supposed belief is related to Prof. Hollnagel's strawman argument about engineers thinking of humans as automatons (noted earlier). The statement simply is not true in engineering or in safety engineering.

Intractable

There is one other term, intractability, upon which Prof. Hollnagel bases his arguments for Safety-II. Intractability is a concept used in computational complexity theory that is used to describe functions for which there exist no efficient algorithms to solve them, that is, the solution on a computer requires an amount of time that is an exponential function of the problem size *n*. It appears that once again, as with bimodality, Prof. Hollnagel borrows a standard engineering term (this time from computer science) and makes up a totally different and unrelated definition.

So let's look at his definition:

> Page 182*: Systems are called intractable if it is difficult or impossible to follow and understand how they function. This typically means that the performance is irregular, that descriptions are complicated in terms of parts and relationships, and that it is difficult to understand the details of how the system works, not least because it changes faster than a description can be completed. Intractable systems are also underspecified, meaning that it is impossible to provide a complete specification of how work should be carried out for a sufficiently large set of situations.*

It is not impossible to understand how engineered (designed) systems, such as cars or airplanes, function. If it were, we would not be able to build and operate them. The way they work does not change faster than a description of how they work. The next phrase about the impossibility of providing "a complete specification of how work should be carried out" once again sounds like he is only talking about humans.

There is an important implication of Prof. Hollnagel's description of systems as being difficult to understand how they work. If this is true of a safety-critical system, then the last thing I want to do is to ask a human to operate it. If engineers cannot understand a system and they designed it, then what chance does a human operator have to understand it and react correctly in real time when something happens differently than expected in the hardware, software, or system environment? Once again, Prof. Hollnagel does not seem to be talking about sociotechnical systems, but only social ones.

> *We can capture these qualities by making a distinction between tractable and intractable systems. […] A system is tractable if the principles of functioning are known, if descriptions are simple and with few details and, most importantly, if it does not change while it is being described. An example could be an assembly line or a suburban railway. Conversely, a system is intractable if the principles of functioning are only partly known (or in extreme cases, completely unknown), if descriptions are elaborate with many details, and if systems change before descriptions can be completed. An example could be emergency management after a natural disaster or, sans comparison, financial markets.*

At last some examples are provided to help understand what he is trying to say. A suburban railway, with today's computer-controlled functions, is not simple (nor are its descriptions) and the descriptions include more than only a few details. It is unclear how Prof. Hollnagel is defining "simple." A system having an elaborate description with many details describes most systems today, although "elaborate" and "many details" are undefined by Prof. Hollnagel. Certainly suburban railways, hopefully, have (or at least should have) elaborate descriptions with many details. His example of an assembly line has nothing to do with product/system safety or, in fact, safety at all.

Let's look at what he calls intractable systems. The first part of his definition—i.e., that principles of functioning are only partly known or, in extreme cases, completely unknown—is irrelevant as engineered, safety-critical systems by definition are designed and the principles of functioning are

known. If the principles of aerodynamics and why the aircraft stays in the air are unknown, then hopefully the passengers all are provided with parachutes. The second part, that the descriptions are elaborate with many details is not a function of the system, but of the description. Descriptions can include many details or few details.

So the difference between intractable and tractable, in the context of this book on safety engineering, must rest on the third part, i.e., that the systems change before descriptions can be completed. Using this definition, Prof. Hollnagel's principle of intractability applies only to natural or non-engineered systems. An example he uses is emergency management. In fact, emergency management after natural disasters is almost always pre-designed. If it isn't, there will not be the supplies, communication devices or channels, trained workers, supply chains, etc. that are necessary to handle an emergency. Most states and governments have emergency management systems that have been designed and can be described. They do not change before descriptions can be completed. It is certainly true that there are emergency situations that can arise without existing detailed plans to handle them or that the plans that do exist are not adequate. So there may need to be flexibility and adaptability in the emergency response. But depending on humans to provide all this flexibility and adaptability without having a pre-designed system that allows humans to be effective in these circumstances would be foolhardy.

While there are aspects of the financial markets that are not designed, many of their functions and controls to avoid "hazards" *are* designed including procedures (sometimes automated) to halt stock market trading when panic buying or selling appears to be happening and myriad rules and regulations to prevent chaos in the markets.

If an emergency does occur in a sociotechnical system, human operators are usually limited in what they can do effectively unless they have predesigned information systems and predesigned controls. It is a mistake for social scientists to treat social systems as totally undesigned and unplanned. The design and provision of existing functions is necessary to provide a system within which humans can be adaptable and flexible. It is not, as claimed by Prof. Hollnagel, impossible to control social systems nor to do risk assessments. The difficulty of doing this and the tools that are necessary to be successful is a useful subject to explore.

> *It follows directly from the definition of tractability that an intractable system also is underspecified. The consequences are both that the predictability is limited and that it is impossible precisely to prescribe what should be done. Underspecification is, of course, only an issue for the human and organisational parts of the system. For the technical parts, in so far as they work by themselves, complete specification is a necessity for their functioning.*

Aha, he *is* only talking about the social parts of systems. However, I have never seen an organizational system that was not specified. In most large companies, management structure, roles, responsibility, accountability, and authority are usually specified in great detail. And we also specify much of the required human behavior. The same is usually true for other types of social systems, such as those required in emergencies. "Underspecification," as used here, is undefined.

> *But this need of a complete technical specification creates a dilemma for socio-technical systems.*

As he just said that technical systems can have a complete specification, then he must be talking only about social systems here, not sociotechnical systems.

> *For such systems the environment cannot be specified completely and it is certainly not constant. In order for the technology to keep working, humans (and organisations) must function as a buffer both between subsystems and between the system and its environment, as something that absorbs excessive variability when there is too much of it and provides variability when there is too little. The problem can in some cases be solved by decoupling parts of the system, or by decomposing it. But for an increasing number of systems this solution is not possible.*

This is the opposite of a sociotechnical or systems approach. In a systems approach, the human operators are not viewed as "buffers" between the social and technical components, but are a part of the overall sociotechnical system that is required for it to function together as a whole to achieve the goals. In fact, this decomposition into three parts that Prof. Hollnagel uses, i.e., the organizational, human operator, and technical, is, in my experience, why we have problems today. A systemic or holistic approach (Safety-III) would think of all of these components as part of one system that must be designed together to achieve its goals, including the need to design for flexibility and adaptability. I don't really understand the relevance of Prof. Hollnagel's discussion of intractability to product/system safety as opposed to purely natural systems.

Some of the confusion in Prof. Hollnagel's book may arise because he mixes up Complexity Theory and System Theory. The difference is explained in Appendix A. I have never seen the term "intractable" systems in systems theory or in complexity theory, but his description (e.g., systems are continually changing and adapting) sounds like he is talking about complexity theory, which has been used to understand and analyze natural systems, such as weather. Systems Theory is the basis for Safety-III and is described later.

Safety Management "Principle"

	Safety-I	Safety-II	System Safety Engineering	Safety-III
Safety Management Principle	Reactive, respond when something happens, or is categorised as an unacceptable risk	Proactive, continuously trying to anticipate developments and events	Concentrates on preventing hazards and accidents but does learn from accidents, incidents, and audits of how the system is performing.	Concentrate on preventing hazards and losses, but learn from accidents, incidents, and audits of how the system is performing.

As stated earlier and should be obvious from the description provided earlier of how various industries manage safety, Safety-I describes something that does not exist. Focusing primarily on reactive measures simply does not accurately describe the safety management practiced by virtually everyone. It doesn't even satisfy Prof. Hollnagel's description. For example:

> Page 58*: Safety-I is reactive and protective because it focuses only on what has gone wrong or could go wrong, and tries to control that, usually by introducing various forms of restriction and compliance.*

"Reactive" is defined in dictionaries as "acting in response to a situation rather than creating or controlling it" or "responding to events after they have happened." Given that every version of system safety engineering described at the beginning of this paper, along with Prof. Hollnagel's own strawman Safety-I, all try to anticipate and control hazards, I don't see how any of them could be defined as reactive—a misstatement he repeats throughout his books. In fact, in the quote from his first book above, he says "*Safety-I is reactive … because it focuses only on what has gone wrong or could go wrong and tries to control that …*" he appears to be describing controlling something that has not yet happened, which violates the definition of reactive.

To be proactive, according to dictionaries, is to be "ready to react or respond to something" or "focusing on eliminating problems before they have a chance to appear." Prof. Hollnagel assigns the adjective of "proactive" to Safety-II:

> Page 139*: From a Safety-II perspective, safety management cannot achieve its stated purposes by responding alone, since that will only correct what has happened. Safety management must instead be proactive, so that adjustments are made before something happens, which will therefore affect how it happens, or even prevent something from happening.*

But Safety-II, in fact, is not proactive—it ignores hazards and just tries to make things "go right." In truth, "proactive" is a better description for the strawman Safety-I and for how safety is actually treated in safety engineering than as a description for Safety-II.

The first goal in safety engineering is to prevent something from happening. But we cannot count on that, so safety engineers (and engineers in general) design to mitigate or control hazards that cannot be eliminated. Even workplace safety tries to anticipate what could "go wrong" and provides procedures, guards, and PPE to prevent accidents.

The social parts of organizations, including safety management, also do this—at least those organizations that survive and flourish. Managers do not wait until a catastrophe occurs before trying to prevent it. For example, competent managers do not wait until the company goes bankrupt before trying to prevent bankruptcy. But if an accident or incident does occur, then something went terribly wrong in our planning and we need to learn what that is from an investigation of the events. As George Santayana warns, those who do not learn from the past are doomed to repeat it. Engineers, including safety engineers, and others are not stupid enough to practice Safety-I. Prof. Hollnagel seems yet again to be arguing against a strawman here.

> Page 139*: For proactive safety management to work, it is necessary to foresee what can happen with acceptable certainty and to have the appropriate means (people and resources) to do something about it.*

This is a description of the primary principle of safety engineering for at least 100 years. We "foresee what can happen" using hazard analysis. Prof. Hollnagel briefly mentions hazard analysis techniques in his book, but apparently does not understand how they are used.

> Page 139: *This understanding can be developed by looking for patterns and relationships across events rather than for causes of individual events.*

We have much more sophisticated ways of understanding hazards and why they occur than randomly looking for patterns. In fact, if we only look for patterns in events, then we are assuming that the world will not change in the future and that only the things that happened in the past will happen in the future. This would be a very naïve way for safety engineers to operate. We identify hazards and then identify the specific scenarios that can lead to accidents in the system being analyzed.

Throughout his books Prof. Hollnagel argues that we should be spending more time making things "go right." In fact, that is exactly what engineers have always spent most of their time on. Aeronautical engineers do not randomly throw pieces of metal together, toss the result into the air, see if it crashes, and then learn from the accident. We design aircraft to stay aloft. It would make no sense not to do this. But if it does not stay aloft, we investigate accidents to determine where our assumptions about how to make things "go right" were wrong. And we also try to design aircraft to be safe if something does go wrong, i.e., operates differently or encounters environments that were not anticipated.

In reality (not in Prof. Hollnagel's strawman), much more effort is put into proactive efforts, than reactive. Even in workplace safety, most of the effort is spent in proactive safety, particularly in the last 50 years. At least ninety percent of engineering is doing exactly what he suggests, but he seems to think does not already happen.

As Prof. Hollnagel himself points out throughout the book, the system and its environment are always changing and we must have a way of coping with these changes. Sophisticated hazard analysis techniques do this and look for relationships across events (through modeling and analysis) rather than

just for causes of individual events. Once again, Prof. Hollnagel is suggesting that we do what we already do, but suggests a less effective way to do it.

Actually, this suggestion of looking for patterns is surprising given that Prof. Hollnagel argues against databases of accidents/incidents.

Investigation/Reporting Databases:

One of the most frustrating parts of Prof. Hollnagel's books are that some of the most effective aspects of safety engineering in the past are denigrated and dismissed. One of these is incident reporting. Prof. Hollnagel writes:

> Pg. 173-174: *One consequence of looking for the devil in the details is the need for large amounts of data about accidents and incidents, typically organized into large databases.*
>
> *[…]*
>
> *One practical consequence of using Safety-II as a basis for safety management is therefore that it reduces the need for such large databases. When we look at things that go wrong and collect accident reports, each incident is treated as a single and special case and must be recorded accordingly. Databases for incidents, for instance in aviation or in health care, are therefore huge and easily run into several hundred thousand reports*.

The major problem in this quote and throughout both of his books is that generalizations are made about how safety is managed in many industries by extrapolating from the limitations in one particular industry and sometimes only in one country and then concluding the same is true for all safety-related industries everywhere. The example he gives here is patient safety reports in Denmark. Many similar problems also exist in healthcare reporting systems in the U.S. But to conclude that databases of accidents or incidents is less useful than looking at what "goes right" is a ridiculous conclusion, even before we consider the problems in Safety-II. An obvious alternative might be to improve the databases of accidents and incidents in healthcare. One of my graduate students is looking at this problem right now and several previous students have also written theses on it. But the same limitations are not true for accident/incident databases in other industries.

Let's examine aviation, as Prof. Hollnagel specifically mentions it in the above quote. There have not been "hundreds of thousands" of aviation accidents. According to the Bureau of Aviation Accidents, including all aircraft capable of carrying more than 6 passengers and excluding helicopters, balloons and combat aircraft, there have been a little over 10,000 accidents since 1970 (the past 50 years). When looking only at scheduled commercial passenger aircraft, the number is much lower. The NTSB aviation accident database contains information about aviation accidents and selected incidents within the United States, its territories and possessions, and in international waters. There are 1574 reports for 2019. The vast majority of these involve general aviation incidents. In 2016, there were 106 accidents in general aviation, 17 of which were fatal accidents with a total of 29 fatalities. In reality, the existing databases on aviation accidents do not contain hundreds of thousands of reports, they are easily used, and they provide important information about how to reduce accidents.

Prof. Hollnagel mixes up incidents and accidents in his quote, so let's look more carefully at incident reporting databases. One of the first incident reporting systems in commercial aviation, the Aviation Safety Reporting System (ASRS) was created in 1976. The ASRS collects voluntarily submitted aviation safety incident/situation reports from pilots, air traffic controllers and others and uses the information to create reports about identified system deficiencies, to issue alert messages, and to disseminate vital information to the aviation community. The collected information is made available to the public and is used by the FAA, NASA and other organizations working in research and flight safety. These reports provide an exceptionally rich source of information for policy development, engineering design, and human factors research.

The first mission of the ASRS is to identify any aviation hazards in reports and flag that information for immediate action. When such hazards are identified, an alerting message is issued to the appropriate FAA office or aviation authority. Their second mission is to classify reports and diagnose the causes underlying each reported event. This analysis has led to changes in cockpit design, for example, to reduce similar human errors.

In fact, the ASRS has been so successful that the FAA, individual airlines, and others have created similar databases and use them to improve safety. These databases provide an important piece in understanding why aviation safety efforts have been so successful in reducing accidents to a very low level.

Prof. Hollnagel makes the following claim (without any evidence, of course)

> Page 174: *Fortunately, it is unnecessary to accumulate hundreds of thousands of individual cases when we look for how things go right. Instead, it is sufficient to develop a description of the daily activity and its expected variability, which means one generic case instead of many specific ones.*

The first part, a description of daily activities sounds like the usual job description, which almost every industry already has. Expected variability is much more difficult to define. Lots of things can go wrong on an aircraft, for example. If the variability described only includes behavior in the normal, expected circumstances, then it might be possible to produce, although I'm not sure exactly what use it would serve. The problem here, again, seems to be that Prof. Hollnagel seems to think that the only components in systems today are human operators.

Expected variability may differ greatly for the same reasons that "work-as-done" differs from "work-as-imagined." The problem is that real-life does not always match what is expected. Serious accidents occur when the unexpected happens and humans often must compensate, under usually extremely stressful circumstances, for what was not foreseen. Describing daily activity is not the same as describing unforeseen circumstances. We cannot describe what we cannot predict and accidents usually involve, again, what engineers call the "unknown unknowns."

In addition, without any examples, I am skeptical that all expected human variability is possible to describe because there are just too many ways to do most jobs and there is no way to determine whether the job/task descriptions actually are safe under all, probably unexpected and unknown, conditions that can occur in a complex system. We already spend a lot of time teaching pilots how to respond in their daily activities. That's not the problem in accidents unless we assume that the accident is totally the fault of the pilot who did not respond with "expected variability." Once again, this approach seems to be a way to blame human operators for all accidents, i.e., they did not respond within the range of expected variability for their tasks.

As for specifying all types of variability, I have investigated many accidents where the investigators were very puzzled about why the human operators did what they did, i.e., varied the way they did their normal job. Trying to predict all of these beforehand would have been impossible. After the fact, it usually made sense, but before the accident it would have been very hard to predict and specify. One of the reasons that hazard analysis uses a backward approach in system safety, i.e., start from the hazards and identify scenarios that can lead to them, is that it is impractical to look at all possible scenarios for operation of a complex system and determine if any of them lead to a hazard.[11] Even with simply looking at human behavior, as Prof. Hollnagel eliminates any consideration of engineering in his writing, it does not seem possible to perform forward analysis and identify all possible behavioral variation as safe or unsafe when jobs are non-trivial. It would be even more difficult to find all the ways that things go "right."

[11] FMECA (Failure Modes and Effects Criticality Analysis) does attempt to do this. But FMECA is virtually impossible to do in a large, complex system. Usually only parts are done or high levels of abstraction are used and with enormous expense. Starting from hazards and identifying the paths to them is much more efficient and practical.

It is curious that Prof. Hollnagel hand waves at the end of his books about doing both Safety-I and Safety-II, but at the same time claims it is too expensive to do "Safety-I." Doing both would then be prohibitively expensive. In fact, decades of experience has shown the great value of accident investigation and incident reporting databases.

> Page 174: *As humans increasingly came to be seen as the root cause of accidents … typical figures being in the order of 80-90 percent, safety came to depend on people reporting on what went wrong.*

This statement by Prof. Hollnagel is again a vast overgeneralization and another strawman argument. Safety does not depend, in any of the industries in which I work, on people reporting what went wrong. Safety is handled in a large number of other ways, as described earlier. But ignoring people's reports of what went "wrong" seems highly irresponsible because it removes an important tool for preventing accidents before they occur. The self-generated reports in the ASRS provide information about behavior of humans and design of systems we would ordinarily never learn about until after a serious accident occurred.

With incident reporting databases, we have been able to find patterns of unsafe or unexpected human operator behavior that are almost always related to system design flaws, such as the design of instruments and controls or procedures, so we can fix them before a plane crashes. These are exactly the patterns that Prof. Hollnagel says we should identify in his supposedly new Safety-II proactive approach—which, in fact, is what people have done for decades. I don't know how those patterns will be found without relying in part on incident and accident databases. In fact, self-reporting with no fear of consequences (part of "just culture") has been remarkably successful, even in healthcare.

But as an engineer, one of the most frustrating parts of Prof. Hollnagel's books is his misunderstanding of the role of learning from failure in engineering.

Learning from Failure in Engineering

Prof. Hollnagel writes:

> Page 173: *One practical consequence of using Safety-II as a basis for safety management is therefore that it reduces the need for such large [accident and incident] databases.*

This statement is only true if one can learn from looking at what goes right and less from what goes wrong. In fact, in engineering, almost all our learning comes from what goes wrong. We learn very little from what goes right. This fact is widely known and commented on in engineering. Some examples:

> Michael Griffin [former head of NASA] in *"System Engineering and the 'Two Cultures' of Engineering"* writes:
> *This brings us to the role of failure in engineering design. Regardless of the sophistication of the analytical methods brought to bear, they are applied to a theoretical model of a device operating in a theoretical model of the real world. The model is not reality, and the differences produce opportunities for the real device to fail to operate as intended in the real environment.*
>
> *…*
>
> *What is of interest in many of the highly public failures which have occurred in large scale systems over the years are not those instances in which something known to have been needed was simply omitted, or those in which a piece-part simply fails. While significant, such cases are relatively easy to understand and correct. What is of interest are those cases, all too many, in which everything thought to be necessary to success was done and yet, in the end, the system did not perform as intended; in a word, it failed.*

In fact, looking only at what was intended, i.e., "success" or "what goes right" to use Prof. Hollnagel's words, will have no impact on the types of accidents that Griffin is describing. It is frequently true that

unintended, unanticipated, and possibly unanticipatible interactions between system elements that were thought to be isolated are the root cause of a problem. It is hard to know how a system can be designed to be resilient in the face of "unknown unknowns". Expecting the human operators to be "resilient" in such circumstances, where they may not even have been provided with the tools to fix the problem or the required information and knowledge to do so is putting an unreasonable burden on them. And it leads to concluding that the cause was inadequate human operator "resilience," the common argument used to assign blame to human operators.

Henry Petroski, a professor of civil engineering at Duke University, has written many thoughtful essays and books on the role of failure in engineering [for example, Petroski, 1992; Petroski, 2006]. While drawn largely from examples in civil engineering, the principles are equally applicable to any engineered system. A key theme of Petroski's writings is that a detailed understanding of the manner in which a given design fails in use allows iterative improvement of later designs. Here are some quotes from his books:

- *"No one wants to learn by mistakes, but we cannot learn enough from successes to go beyond the state of the art."*
- *"Past successes may be inspirational and encouraging, but they are not by themselves reliable indicators of or guides to future success. The most efficacious changes in any system are informed not by successes but by failures. The surest way for the designer of any system to achieve success is to recognize and correct the flaws of predecessor systems, whether they be in building codes or in banking policies or in bridges."*
- *"Failure is central to engineering. Every single calculation that an engineer makes is a failure calculation. Successful engineering is all about understanding how things break or fail."*
- *"A failed structure provides a counterexample to a hypothesis and shows us incontrovertibly what cannot be done, while a structure that stands without incident often conceals whatever lessons or caveats it might hold for the next generation of engineers."*

In the past, structures were designed and built according to "rules of thumb" that were derived from prior successes and failures and the designs passed down from master to apprentice. But today, we have a sophisticated theory of the strength of materials and of structural design generally; we know how to design optimal structures for a wide range of user-specified optimality criteria. But structures occasionally fall down. Rarely do those accidents require making changes to the theory. More often, the failures result from introduction of new materials, new uses for standard materials, changes in construction practices, changes in the environment, etc. If we look only at the millions of times that structures do not fall down, we learn nothing. Knowledge advances when the assumptions we make about systems are violated and a loss occurs.

After sometimes billions of simulations of automated car designs, a car may be put on the road and an accident occurs. This phenomenon can be explained by the fact that accidents occur when the assumptions we make about the way the system will work or about its environment turn out to be false. Those same assumptions are used in testing and simulation. We cannot determine that the assumptions are wrong until they fail, usually in odd cases that nobody anticipated. Sometimes the assumptions originally were right but the world has changed. More often we just have not yet come across the unique and unusual failure conditions. We get more confidence that what we are doing is right the longer it takes to come across these conditions. If it takes a while, we become more convinced that our assumptions are right.

In fact, most major accidents result from the overconfidence that comes with success. We believe that risk is decreasing or has decreased when, in fact, the risk is the same as it always was. This is the folly of trying to learn from "success" or lack of bad consequences. Continuing success only gives us

more confidence that what we are doing is the safe thing to do. Prof. Hollnagel points out many times in his books that accidents are rare. It is for this reason that we have to learn from the failures and accidents that *do* occur. It is from failure that engineers learn how to be successful in the future.

Is this true only for engineering and not for human performance? Here is one final quote. It is from Russell Ackoff, a professor of Management Science at the Wharton School, University of Pennsylvania, and one of the great systems thinkers of our time:

> *All learning ultimately derives from mistakes. When we do something right, we already know how to do it; the most we get out of it is confirmation of our rightness.*

Ackoff was not writing about the technical parts of systems here, but the human parts. Some of the most important lessons from the investigation of major accidents such as Three Mile Island, Deepwater Horizon, and Fukushima-Daiichi, have involved design errors and unsafe behavior in the social parts of these systems, including human operators and oversight agencies.

Let's take a simple example. Suppose that my husband drives our car frequently and always fills the gas tank before returning home. As a result, when I drive the car, I never bother to check the gas level. One day, something unusual happens and my husband does not fill the gas tank, and I am stranded far from a gas station. What did I learn from the times that things went right? In fact, I foolishly learned that I did not need to need to check the gas level.

Consider Figure 5. A designer creates a design using ideals and averages for materials, etc. When the system is constructed, the actual materials used may differ from the averages assumed in the design specification. Over time, the system will evolve and change in significant ways from that originally created. The designer also provides operational procedures and training materials for the operators based on the original design specification or model.

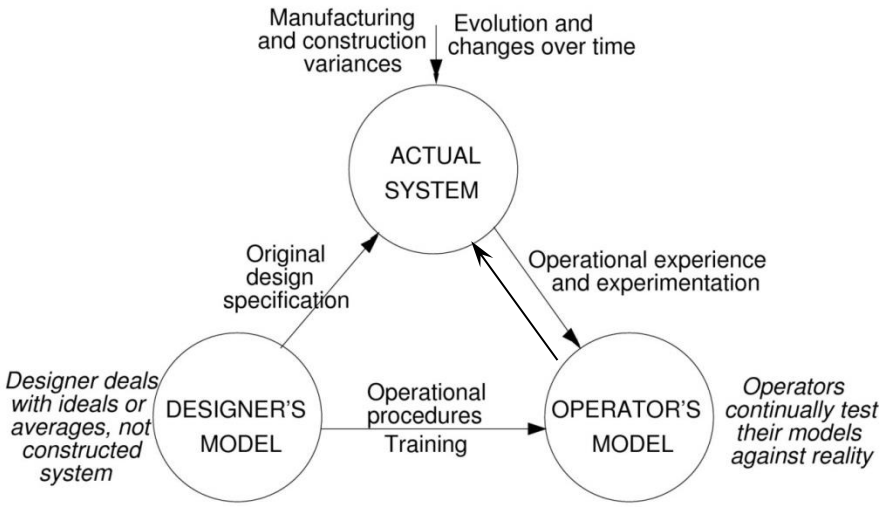


Figure 5: Operators learn from crossing the boundaries of safe behavior

The operators start with a mental model of the system that conforms with the original design specification, but they must operate the system as it exists at the time and not according to an ideal conception of it in the designer's mental model, which has been incorporated in the training materials for the operators. Operators then must continually test their models against reality to operate the system efficiently and safely. I agree with Prof. Hollnagel that operators must learn from operational

experimentation on systems, but if the operator's hypotheses are incorrect, then that experimentation can lead to accidents.

Because accidents are rare, as are the conditions that lead to them, operators may assume that something is safe because nothing untoward happens for the first hundred times they do something. If the conditions change, however, the next time can lead to an accident. In today's highly complex systems, it is not possible for the operators to understand the system design in depth and to assure that their experience in the past applies to the present state. Often, the only way to identify where the boundaries of safe operation is to go past them. An accident does not always result every time they do that. How do they know they have passed the safe boundaries when no accident occurs? Have they learned to do unsafe things?

Jens Rasmussen has suggested that, in the search for efficiency and productivity, we tend to migrate to the boundaries of safe behavior. We don't know where those boundaries are, however, until we go over them and have incidents or accidents [Rasmussen, 1997].

That doesn't mean that we cannot learn anything when "things go right," i.e., when nothing untoward happens, but we can also learn the wrong things at this time unless we have complete understanding of how the system works and the current system state. That is rarely the case.

There is one other strange suggestion by Prof. Hollnagel regarding accident investigation where he claims that in order to optimize learning, we should use frequency rather than severity to determine what to investigate.

> "*Small but frequent events are furthermore easier to understand and easier to manage. It there makes better sense to base learning on those than on rare events with severe outcomes….*"

The problem with the logic here is that he bases his argument on Heinrich's pyramid, which was created in the 1930's for workplace accidents only. Heinrich's hypothesis about ratios between serious accidents and less serious incidents does not apply to complex systems and product safety [Amalberti, 2001]. In complex, engineered systems, accidents may result (and often do) from very different things than incidents, i.e., the causality may not be related. So looking only at frequent events with minor consequence may tell us nothing about more serious events.

Accident Causality and Causality Models

	Safety-I	Safety-II	System Safety Engineering	Safety-III
Explanations of accidents	Accidents are caused by failures and malfunctions. The purpose of an investigation is to identify causes and contributory factors.	Things basically happen in the same way, regardless of the outcome. The purpose of an investigation is to understand how things usually go right as a basis for explaining how things go wrong.	Accidents are caused by linear chains of failure events. The purpose of investigation is to identify the chain of events and the root cause.	Accidents are caused by inadequate control over hazards. Linear causality is not assumed. There is no such thing as a root cause. The entire socio-technical system must be designed to prevent hazards; the goal of investigation is to identify why the safety control structure did not prevent the loss.

In this one case, Prof. Hollnagel does describe Safety-I in his table as what is common practice. But in this case, the statements about Safety-II are fairly strange, perhaps again because of his use of imprecise language. Safety-III, as will be seen, differs significantly from what is usually done today.

Things basically happen in the same way, regardless of the outcome.

What "things"? Everything happens in the same way? What could "happen in the same way" possibly mean? If I forget to turn off a faucet, it can lead to a flood in my kitchen but it will not lead to my car having a flat tire or even to my car operating correctly. He must mean that the general set of causes of things "going wrong" is the same as the general set of causes for things "going right"? Doesn't the outcome of an action depend on what has happened prior to the outcome or on the current conditions? Why are no examples included for this seemingly odd statement?

There is a figure (7.3, page 137 and reproduced below) in his first book that appears to explain his statement by implying that everything occurs because of performance variability (everyday work). Notice that this is just a variant of his Figure 3.2 reproduced above, which did not make much sense because the terminology was undefined and not the way that safety engineering thinks of causation (see Figure 4). It makes no more sense in this variant of Figure 3.2. But let's assume it does.

Of course human performance varies and some particular instances of that variance will lead to accidents and other instances will not. The difference (as discussed in more depth below under accident causality models) is that "Safety-I" describes an accident as resulting from malfunctions or errors during "everyday work" and leading to a hazard while "Safety-II describes accidents as resulting from performance varying and crossing safe boundaries—which is the definition of an error or malfunction—leading to a hazard. In his Safety-II definition, behavior varying beyond safe boundaries and thus leading to an accident is not the same reason that accidents do not occur, which is that behavior remains within safe boundaries and hazards do not occur. These are not "things happening in the same way." Both in his description of Safety-I and Safety –II, the cause of accidents occur in different ways than not having accidents, but both occur during "everyday work." What is not everyday work?

The primary difference between his Safety-I view of causality and Safety-II is that causality in Safety-I seems to be general and apply beyond human operators. Also, in Safety-I, there seems to be an implication that all functioning (everyday work) follows "work as imagined" or specified and any deviation from that will lead to an accident. This is simply another strawman.

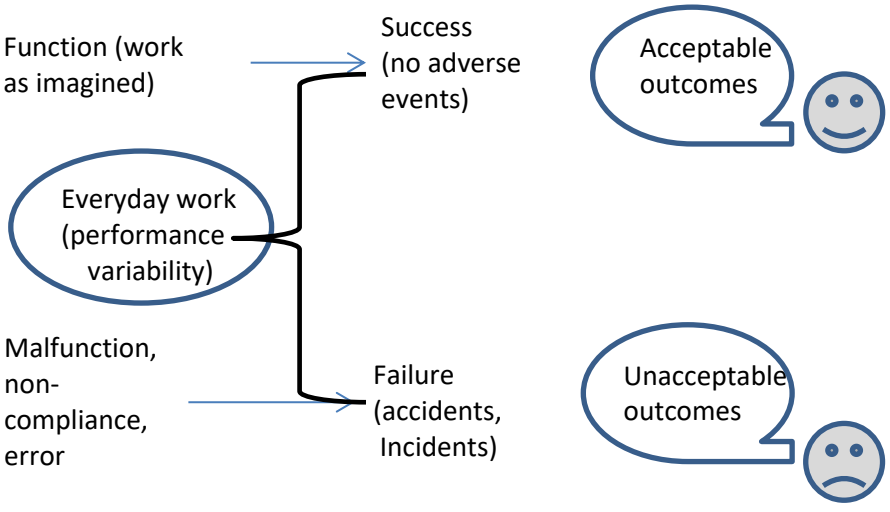


Figure 6: (Hollnagel Figure 7.3 on Page 137): "The Safety-II view of failures and successes"

The rest of his description of the Safety-II explanation for accidents does not make much more sense.

> *The purpose of an investigation is to understand how things usually go right as a basis for explaining how things go wrong.*

I have participated in a lot of accident investigations. I can't imagine why I would spend my time understanding how things usually "go right." I know that before the accident occurred. What I don't know is why it "went wrong." I know why engines operate correctly, why the airplane maintains lift and stays in the air, and why my brakes usually work to stop my car when I press on the brake pedal. I designed them to do those things. Wasting my time on things I already know is not productive. When an accident occurs, the system (including the human operators) did not work the way I expected it to work. Why did the engine stop working, the aircraft lose lift, or the brakes fail? That's what I need to understand.

My examples in the previous paragraph are all in the engineering realm. Perhaps Prof. Hollnagel is assuming all accidents are caused by human operators, as he does throughout his two books on Safety-II. If the pilot usually points the nose of the aircraft down when the aircraft is about to stall, then things "go right." I already know that. What I need to know when an accident occurs is why the pilot did <u>not</u> do the safe thing, i.e., why did it make sense to the pilot to do what led to the accident or not to do what would have prevented the loss? There has to be a reason, and knowing that reason will help me understand why the accident occurred and what to change to prevent a similar one in the future, which is the goal of accident investigation.

When the driver does not activate the brakes in time to prevent hitting the car ahead, I need to know why. Perhaps the driver was daydreaming and not paying attention to traffic. Investigating why the driver usually does activate the brakes (why things usually "go right) could include he was paying attention, the car in front did not stop abruptly, the car ahead was clearly visible due to the weather conditions at the time, or the driver was not paying attention but there was nobody in front at the time or by chance the driver's attention was regained in time to activate the brakes successfully. There are a lot of reasons why drivers do <u>*not*</u> crash into a stopped car in front of them (things go right), but I don't understand how that helps me explain why this accident occurred except to say that those conditions did not hold in this case.

Even more important, Prof. Hollnagel is suggesting that accidents are always caused or prevented by human variability. What about the technical parts of the system? What about the aspects of the system or its environment upon which the human operators have no control? The flaw in the logic here is that all accidents are not caused by or prevented by operator behavior. In an accident investigation, we already have the information about why it usually "goes right," i.e., because we designed it to "go right." Instead, we need to spend our time identifying our flawed assumptions that led to an unsafe system design and flawed assumptions about how humans would behave within that system design.

> Page 163: *[Using Safety-II for investigation] is done by constructing an account of everyday, successful performance and then looking for how performance variability, alone or in combination, could lead to loss of control. … Asking for what went <u>wrong</u> leads to a search for errors, failures, and malfunctions. Asking for <u>what did not go right</u> creates a need to understand how work normally takes place and how things go right.*

Prof. Hollnagel is playing with words here again and his vague definitions of "right" and "wrong." Asking "what went wrong" in an accident, using the braking example again, is to ask why the driver did not press the brakes in time to prevent the collision. Asking "what did not go right" involves asking why the driver did not press the brakes in time to prevent the collision. They are identical.

We already know why drivers normally press the brakes and stop the car before a collision occurs. We need to know why that did not occur in this case, i.e., why the behavior varied from the usual behavior or, in other words, varied beyond the safe boundaries. Things that "go right" here happen

because the driver presses the brakes in time to prevent a collision. Things that "go wrong" here happen because the driver does not press the brakes in time to prevent a collision. They do not happen in the same way, as claimed.

If I am misinterpreting what he is saying, then that might have been avoided by providing some examples or using carefully defined terminology.

To understand columns three and four in the table above, i.e., how safety engineering normally explains accidents and how that might be improved in Safety-III, more information about the basic concept of causality and models of accident causality is needed. Once again, Prof. Hollnagel's definitions and concepts differ from the standard ones.

Causality in General

Prof. Hollnagel makes many unusual statements about causality throughout the book, all of which seem suspect (at the least) and none of which, of course, are explained beyond handwaving. Almost no examples are provided. These statements include:

> Page 52*: "Since the cause usually can be associated with a specific component..."*

I have been involved in accident investigation and reading hundreds of accident reports for 40 years. In my experience, it is never true that the cause is associated only with a specific system component, except for the fact that accidents are usually blamed on the human operators. That blame, however, does not mean that all accidents actually are caused by human operators. And usually other causes are acknowledged as at least contributing to the loss.

Accident causality is always complex. In a well investigated accident, there are almost always physical, technical, operational, social, managerial and governmental contributors. In the Bhopal accident, there were technical factors, physical design errors in the plant as well as operator errors, management negligence, regulatory deficiencies, international agricultural policies, and more involved. Taking only one of these perspectives or simplifying the explanation or cause for the accident does not allow us to learn as much as possible from an accident and apply these lessons to the future.

Other strange statements he makes about causation in general include:

> Page 105: "*In a growing number of cases we must stop taking for granted that we can find the cause when something has happened—or even that we can come up with a plausible explanation"*
>
> Page 104: "*as the world …becomes more complicated, it is no longer reasonable to assume that we can understand the causal relationships between actions and outcomes, or even that they can be described in causal terms."*
>
> Page 61-62: He labels as a myth (and thus untrue)*: "Outcomes can be understood as effects that follow from prior causes"*

These statements differ from standard definitions and assumptions about causality, but are not supported by any evidence, examples, or scientific reasoning.

Western philosophers have debated the notion of causality for centuries. The standard models of causality used in safety are consistent with the definition by John Stuart Mill (1806-1873) that a cause is a set of *sufficient conditions*. "The cause is the sum total of the conditions positive and negative, taken together, the whole of the contingencies of every description, which being realized, the consequences invariably follows" [Mill, 1843].

As an example, combustion requires a flammable material, a source of ignition, and oxygen. Each of these conditions is necessary, but only together are they sufficient. The cause, then, is all three conditions, not one of them alone. The distinction between *necessary* and *sufficient* is important [Lewycky, 1987]. An event may be caused by five conditions, but conditions 1 and 2 together may be

able to produce the effect (represented by an AND in a fault tree), while conditions 3, 4, and 5 may also be able to do so (represented by another AND in a fault tree. The two possibilities would be represented by an OR in the fault tree. Thus, there are two causes (set of conditions sufficient for the event to occur); both of the causes have a set of necessary conditions. The standard chain-of-events model used in safety (see below) is based on this standard definition of causality.

Note that, in this definition, simple association is not enough. A condition may precede another event without causing it. For Mill and those who use his definition, a regular association of events is causal only if it is "unconditional"—that is, only if its occurrence does not depend on the presence of further factors such that, given their presence, the effect would occur even if its putative cause was not present. Thus, a condition or event may precede another event without causing it.

Furthermore, in this definition of causality, a condition is not considered to cause an event without the event happening every time the condition holds. Drunk driving, for example, is said to cause accidents, but being drunk while driving a car does not always lead to an accident. And a car accident does not only result from a drunk driver. The same type of condition holds between smoking and lung cancer. Therefore, using this standard definition of causality, drunk driving does not "cause" automobile accidents and smoking does not cause lung cancer. The Tobacco Institute made this argument for decades. This does not mean, however, that there is no causal connection, but the conditions may be indirect or depend on additional factors being true.

Both workplace safety and product/system safety have used this traditional conception of causality. In fact, it is common to almost all western society. Without some model of causality, we would be faced with a totally random world, with few tools to assist in engineering or in preventing accidents. But that does not mean that we are limited to the model of causality defined two hundred years ago.

Given this background, let's examine some of the arguments that Prof. Hollnagel makes about causality. He describes something he calls "the Causality Credo" that he implies is a widespread but false belief. There are, of course, no examples and no citations. Once again, he creates a strawman (using vague and undefined terminology) without any references to people who might believe it and then easily demolishes it.

> Page 87*: Causality Credo: Belief that all accidents are preventable---all accidents have causes, and causes can be found and eliminated*

His argument to support this "credo" starts with:

> *The reason for adverse outcomes (accidents, incidents) is that something has gone wrong (cf. Figure 3.2). Similarly, the reason for successful outcomes is that everything worked as it should, although this is rarely considered.*

The problems with Figure 3.2 were covered previously. What does "worked as it should" mean? What about "something has gone wrong"? This seems like a truism or circular argument. If I have an accident, then by definition something has gone wrong as I did not intend to have an accident. If everything works as it should, then by definition, everything worked as it should and I did not have an accident. Doing something is not right or wrong except with respect to specified outcomes. As discussed previously, engineers would never use such vague, undefined language.

Why would anyone "rarely consider" that the reason for successful outcomes is that everything worked as it should? Is there some data or some examples where people did not think that successful outcomes occurred because everything "worked as it should"? If my brakes stop my car when I activate them, why would I not consider that they worked as they should? Why would such thinking be rare? In fact, the statement appears to be a truism. If something has gone wrong, then everything has *not* worked as it "should" simply by the definition of "should." And if everything worked as it should, then nothing has gone wrong. What other possibility is there? Again, the loose and undefined language makes interpreting the points being made almost impossible but does allow for fallacious reasoning.

Starting from this basis, Prof. Hollnagel makes the following argument:

1. *Since adverse outcomes have causes, it must be possible to find these causes provided enough evidence is collected. Once the causes have been found, they can be eliminated, encapsulated, or otherwise neutralized.*

This is not engineering language. I have no idea what "encapsulating and neutralizing" something means except perhaps for toxic substances. Identifying causes does not necessarily mean something can be done about them. Or that it would be cost-effective to do so. It might be better to simply not use such a system in critical applications.

Doing so will reduce the number of things that go wrong, and hence improve safety.

Reducing the number of things that "go wrong" will not necessarily improve safety, where safety is defined in terms of the number of accidents/hazards or their impact. Design can reduce the number of things that can go wrong, but not necessarily the things that do go wrong. Once again, the arguments do not make any sense from an engineering perspective. And, once again, what does "go wrong" mean?

Since all adverse outcomes have a cause (or causes?) and since all causes can be found, it follows that all accidents can be prevented.

Even if it were true that all causes can be found, it is not true that all causes and thus accidents can be prevented. There is another strawman argument here. And how many people really believe that all accidents can be prevented? Certainly not engineers. But Prof. Hollnagel is arguing here that this is what engineers believe. On the basis of what evidence?

2. *This is the vision of zero-accidents or zero harm that many companies covet.*

The goal of "zero accidents" is a phrase commonly used in workplace safety; it is not used in product/system safety. The goal in system/product safety is to eliminate or mitigate hazards. Of course, not all accidents can be prevented. Doing so may be too expensive or impractical, even if we theoretically could.

There is much misinformation or confusion about engineering in all these points. The causality credo seems to be something that Prof. Hollnagel has made up to try to further his arguments about Safety-II. Or perhaps he just does not understand engineering.

Here is an alternative argument about causality that most engineers (and probably non-engineers) *could* accept:

1. We can perform hazard analysis in addition to using past experience to identify many causes of accidents.
2. In many cases, but not in all, these causes can be eliminated and controlled.
3. If they cannot be eliminated or controlled, then difficult (and often trans-scientific) decisions must be made about whether the system should be constructed or perhaps used only in limited circumstances where the risk is acceptable to all the stakeholders.
4. Usually a large number of accidents can be prevented but not all.

It is the engineer's responsibility to outline the hazards (risks) but only the stakeholders can determine the fate of the system. Are the tradeoffs worth the risks? Who makes such a decision? Prof. Hollnagel here is confusing the *goal* of engineering with how well that goal can be achieved.

As stated earlier, Prof. Hollnagel's arguments are based on some unusual views of causality:

p. 105: *In a growing number of cases we must stop taking for granted that we can find the cause when something has happened---or even that we can come up with a plausible explanation.*

A statement like this cries out for an example of an accident where we did not find the cause along with some data to show that this phenomenon is increasing (growing). But again, no examples or data are

provided. While Prof. Hollnagel's statement may perhaps be true for natural systems (volcanoes or weather phenomena) it is rarely true for engineered systems---we have the plans and know how they were put together and rarely build something for which there are parts whose behavior we do not understand. And, of course, we should never do this when engineering safety-critical systems.

There probably are some accidents for which causes cannot be found although I cannot think of any. But if the statement is true, then Prof. Hollnagel should have been able to come up with an example, especially if this phenomenon in increasing.

> Page 125: *Safety-I … assumes that adverse outcomes can be explained in terms of cause-effect relationships, either simple or composite and the ontology assumes that fundamentally something can either function or malfunction.*

Not just Safety-I, but most scientists and engineers for several hundred years explain behavior in terms of cause and effect. The rest of the sentence depends on what is the definition of "function" and "malfunction," which, of course, Prof. Hollnagel does not define anywhere. By the normal definition, malfunctioning is the opposite of functioning so the statement is a truism. If a different definition of "malfunction" is intended, it should have been provided. Statements cannot be right or wrong unless the terminology is defined.

That effects follow from causes has long been an assumption of philosophy and the definition of causality. It has nothing to do with any so-called "Safety-I". If it is not true, then the world is random and outcomes do not have causes but occur for no reason. Not to get too philosophical here, events often appear random because perfect information is not available. More information usually lessens the belief that the outcomes are random.

Prof. Hollnagel summarizes his argument about causality by:

> *"as the world …becomes more complicated, it is no longer reasonable to assume that we can understand the causal relationships between actions and outcomes, or even that they can be described in causal terms."*

As a safety engineer, I am not concerned with understanding the behavior of the "world" as a whole, only the behavior of the safety-critical systems that I design. If I cannot understand their behavior in terms of the causality of the hazards, then it would be unethical for me to allow the systems I design to be used where human life is involved.

Perhaps his conclusion stems from his confusion about the term "emergence" discussed earlier. Emergent properties in complex systems are not those for which we do not understand the causal relationships, only that the causality cannot be completely described by looking at the components and not also at their interactions.

Many of the logical problems in the arguments in the book stem from a misunderstanding about models of causality. The engineering of safety-critical systems is based on our models or beliefs about how accidents are caused. So let's look further at models of accident causality, a subset of causality models in general.

Models of Accident Causality

Models of accident causality underlie everything we do to engineer for safety. The way we explain why accidents occur, that is, the model of causality, determines the way we go about preventing them and learning from those that do occur. Most people are not aware they are using an accident causality model, but they are. Basically, models impose patterns on the events we see. Models represent our assumptions about how the world operates. For example, if an underlying assumption is that accidents are caused by human operator error, then the analysis will focus on what the operators did to contribute to or prevent the loss. Such assumptions about the causes of accidents always underlie

engineering for safety, but those doing the analysis may be unaware of any subconscious assumptions they are making.

Prof. Hollnagel states:

> *It is today common to distinguish among the at least three different types of accident models, called the sequential, the epidemiological, and the systemic, respectively.*

This terminology is far from "common." In fact, I have only seen it in Prof. Hollnagel's writing (even before the books on Safety-II) and in a few places that cite him. His definitions do not fit the standard terminology in safety engineering (or epidemiology) for about 50 years. He never does define the "systemic" model anywhere in the two books on Safety-II. The term systemic model is used by others, including myself, to mean an accident model based on Systems Theory, but it is difficult to tell whether he means the same thing without any definition being provided.

Let's examine more standard definitions of accident models, starting with the almost universal model of causality used in engineering, i.e., the model of accidents as resulting from a linear chain-of-failure events. Note that the mathematical definition of linearity and nonlinearity is not being used here. Prof. Hollnagel confuses the two in his books. In mathematics and science, a *nonlinear system* is a system in which the change of the output is not proportional to the change of the input. Nonlinear problems are of interest because most natural systems are inherently nonlinear. Nonlinear dynamical systems, describing changes in variables over time, may appear chaotic, unpredictable, or counterintuitive, contrasting with much simpler linear systems. In contrast, *linear causality* (vs. mathematical linearity) is a conception of causation as following a single, linear direction, i.e. event A causes effect B, where B has no demonstrable effect on A. A synonym for linear causality here might be "sequential causality."

Three causality models are described here: the Linear Chain-of-Failure Events Model (and some subsets such as the Domino Model, the Swiss Cheese Model, and Prof. Hollnagel's Resonance "Model"), the Epidemiological Model, and STAMP (System-Theoretical Accident Model and Processes).

The Linear Chain-of-Failure Events Model

This common model is built on the assumption that accidents are caused by chains of failure events, each failure being the direct consequence of the one before. For example, someone enters the lane in front of your car, you slam on the brakes but are too late in applying them, and therefore you hit the car in front of you. Perhaps in addition, someone was following too close behind you and rear ends your car.

Figure 7 shows an example of applying a simple chain of events model for a tank explosion. Note that the chain can have logical "ANDs" and "ORs" in it—it is still considered a linear chain. In this accident, moisture gets into the tank, which leads to corrosion, which in turn causes weakened metal. The weakened metal along with too high an operating pressure for the condition of the tank leads to a tank rupture, which causes fragments to be projected. The fragments lead to damaged equipment and/or personnel injury.

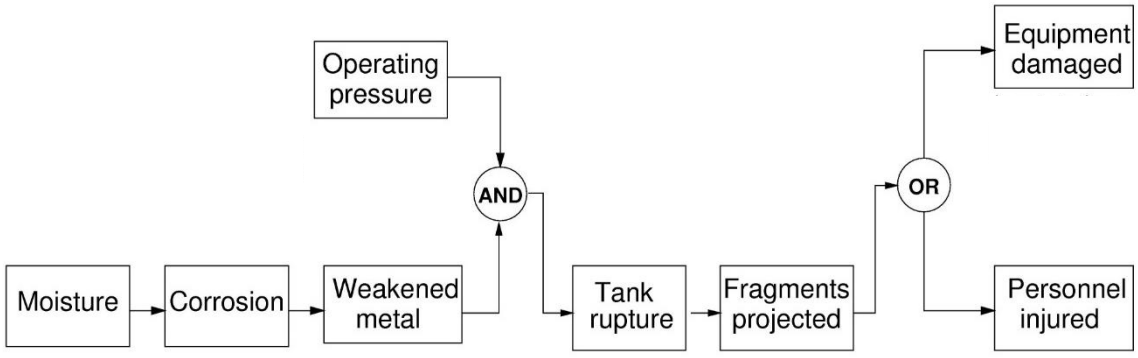


Figure 7: Chain of events model for a tank explosion

Using this model of accident causation, it appears that the simplest way to prevent such an accident is to eliminate one of the events in the chain. An alternative is to put barriers between the events so that the consequences of one event does not lead to the next event in the chain. An example of a barrier in this case is to put a screen around the tank so that in the event of a rupture, the fragments cannot be projected outside a protected area. The use of barriers is common practice in the nuclear power and the process industries that use "defense in depth" to prevent accidents. In such approaches, multiple barriers are provided, where each barrier is used to back up the previous one. For example, protective cladding is put around the nuclear fuel to contain the radiation. In case the cladding is not effective, a shutdown system is used to stop the reaction. A final defense, if everything else fails, is the containment structure surrounding the entire reactor building.

Figure 8 shows an annotated model of the same tank explosion accident chain where possible protection or control activities are added. A wide selection of preventive activities are included in the example: activities or design features that prevent the event and those that prevent propagation (barriers between the events). For example, moisture might be kept out of the tank by using a desiccant or the tank might be coated with stainless steel to prevent corrosion.

There are a few things to note here. The first is that direct causality is assumed, that is, each event leads to the next event in the chain. Also, the preceding event is necessary for the following event to occur, i.e., if moisture does not get into the tank, then corrosion will not occur. That is, the previous event in the chain is both necessary and sufficient to produce the following event, as Mill suggested as the definition of causality 200 years ago.

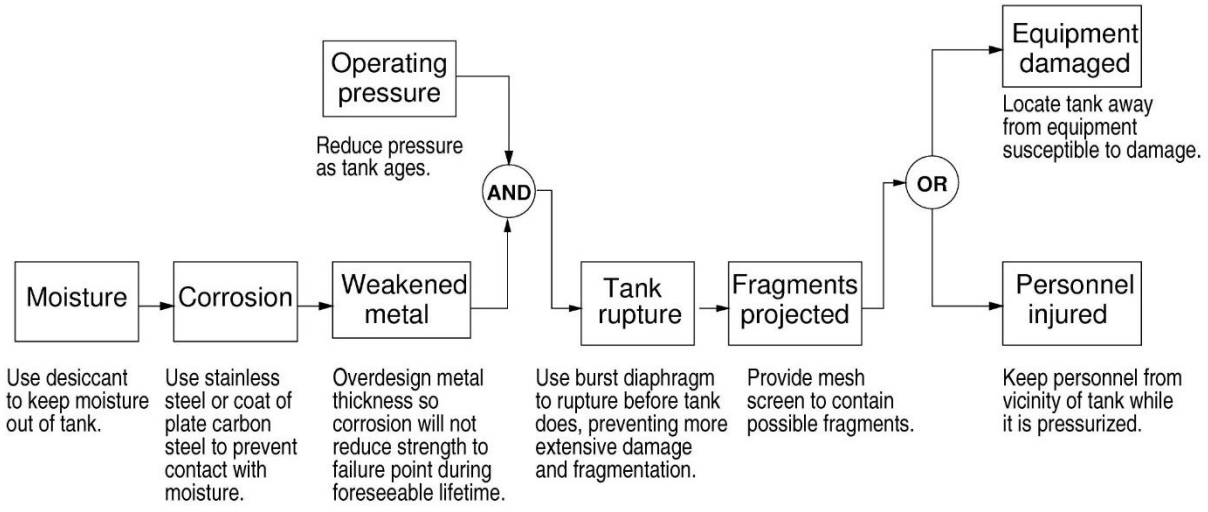


Figure 8: Tank explosion example shown with added protections

Using this model to explain a loss that has occurred, the analysis works backward from the loss event to identify the causal chain. The initial event is called the "root cause." While the event labeled the root cause does not have to be the first one documented in the chain, it usually is. Note that almost always the stopping point is arbitrary and often the root cause is identified as a human operator. In Figure 7 or 8, more previous events could be added, which would then be the "root cause." The search works backward until something is found that is easy to prevent or the search cannot easily go backward any

farther. That event is labeled the root cause. Sometimes politics or liability become involved in the selection of a root cause.

As an example of how events could be added, consider the first event in Figure 7 or 8, which is moisture entering the tank. The moisture must be introduced somehow, and there probably were design features used to prevent water and moisture reaching the tank. Water getting into the tank and the provision of protection devices to prevent this event could be added to the beginning of the chain. Is that failure the root cause? The event that is chosen as the root cause is arbitrary as any of the events in the chain could theoretically be labeled as such or the chain could be followed back farther.

Previous events and other events in the chain could involve management decision making and other social factors, e.g., the regulatory agency approves a flawed design, regulatory procedures are not created that prohibit the specific design flaws, etc. These are all events in the chain of events.

There are no operators in the simple example shown, but usually an operator is selected as the root cause. We might change the example to have an operator opening a valve that allows moisture to get into the tank. Rasmussen suggested that one reason that operators are usually chosen as the root cause is that it is difficult to go backward through a person to an event that causes the operator error [Rasmussen, 1990]. The interface design, for example, is not an event. It is a feature of the system design or the context in which the operator is working. What is the direct cause of the pilot giving an incorrect command to the flight management system? This is one reason why operators are usually blamed for accidents, although there are other reasons. It would be possible to add an event that involved the designer deciding to include a particular feature in the interface or involving a review of the interface design, but I have never seen that done.

Software also is not included in Figure 7 or 8, but we could have, for example, by having software control the burst diaphragm or relief valve in the tank. While it is easy to say "Software does not open the relief valve" (perhaps in a box preceding "Tank Rupture"), it is more difficult to think of a way to protect against this behavior. Only a very small part of real software can be tested within a reasonable amount of time. Software is an abstraction (set of instructions) that cannot fail—it does exactly what it was told to do so the problem must involve a system design or requirements error on the part of the engineers. How does one create simple protections or barriers against that?

Note that the other boxes in the chain might also contain design errors (the design of the tank, for example) but those causes are omitted from the chain of events model because they are not events. I will come back to this later. I have noticed that in most real-world hazard analyses, there usually is not much included about software or the design of the product (e.g., the aircraft) as a causal factor. Human error and physical failures are the primary causes considered. The most common root cause selected is a human error.

In the accident report of the American Airlines B757 crash while landing at Cali, Colombia, there were four causes cited, all of them involving something the flightcrew did or not do. The strangest was the fourth one: "*Failure of the flight crew to revert to basic radio navigation at the time when the FMS[automation]-assisted navigation became confusing and demanded an excessive workload in a critical phase of flight."* The cause was that the flightcrew became confused rather than the cause being that the automation was confusing?

The Domino Model

Heinrich, who created the Domino Model, worked for an insurance company and was primarily interested in workplace safety. He was convinced that humans were the cause of most accidents and, in 1931 created his *Domino Model* to explain this phenomenon (Figure 9).

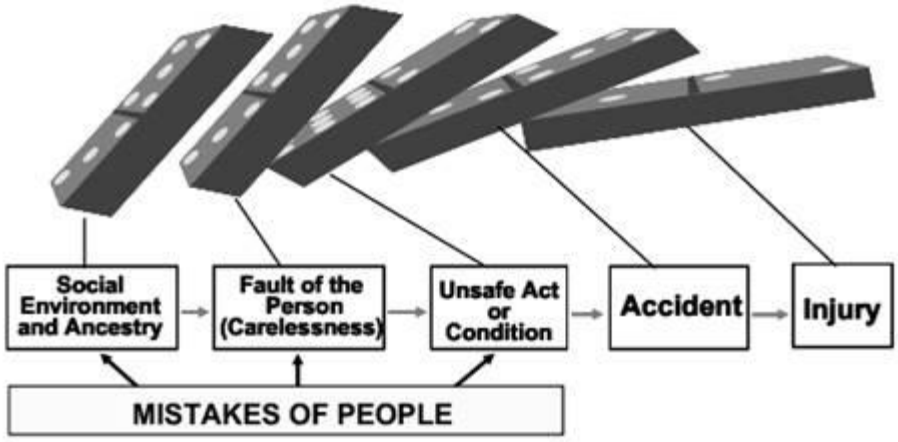


Figure 9: Heinrich's Domino Model of accident causation

Note that this is simply a special or limited case of the chain of events model where the events are depicted as five dominos, with each domino "falling" and causing the next domino to fall instead of using an arrow between square boxes. While unsafe conditions are included in the third domino, they are assumed (as shown) to be the result of "mistakes of people."

Removing any of the dominoes will break the sequence and prevent the injury, but Heinrich argued that the easiest and more effective domino to remove was the third one, representing an unsafe act or condition. Unfortunately, Heinrich was very successful in promoting worker error as the primary cause of accidents in the workplace, and his causality model persists today.

Heinrich's model was extended by others to include more factors. For example, Bird and Loftus in 1976 extended it to include management decisions as a factor in accidents [28]. The modified chain or sequence of events was defined as:

1. Lack of control by management, permitting
2. Basic causes (personal and job factors) that lead to
3. Immediate causes (substandard practices/conditions/errors), which are the proximate cause of
4. An accident or incident, which results in
5. A loss.

In this model, the four major elements of a business operation—people, equipment, materials, and environment—individually or in combination are the factors involved in a particular accident.

Adams, also in 1976, suggested a modified and more general management model [2] that included:

1. Management structures (objectives, organization, operations)
2. Operational errors (management or supervisor behavior)
3. Tactical errors (caused by employee behavior and work conditions)
4. Accident or incident
5. Injury or damage to persons or property.

The commonality among all of these is the linearity of the causal chain and fact that human error is behind most of the events in the chain.

The Swiss Cheese Model

Prof. Hollnagel writes:

> *The simple, linear model was superseded in the 1980s by the epidemiological model, the best known example of which is the Swiss cheese model. The Swiss cheese model represents events in terms of composite linear causality, where adverse outcomes are due to combinations of active failures (or unsafe acts) and latent conditions (hazards). Event analysis thus looks for how degraded barriers or defenses can combine with active (human) failures. Similarly risk analysis focuses on finding the conditions under which combinations and single failures and latent conditions may result in an adverse outcome, where the latent conditions are conceived as degraded barriers or weakened defenses.*[p. 66]

In fact, the Swiss cheese model is the same type of linear causality chain model as the older ones, with, once again, a spotlight and emphasis on worker error. It is just a minor variation (mostly graphical) of the Heinrich, Bird and Loftus, and Adams models. In addition, epidemiological models are something quite different.

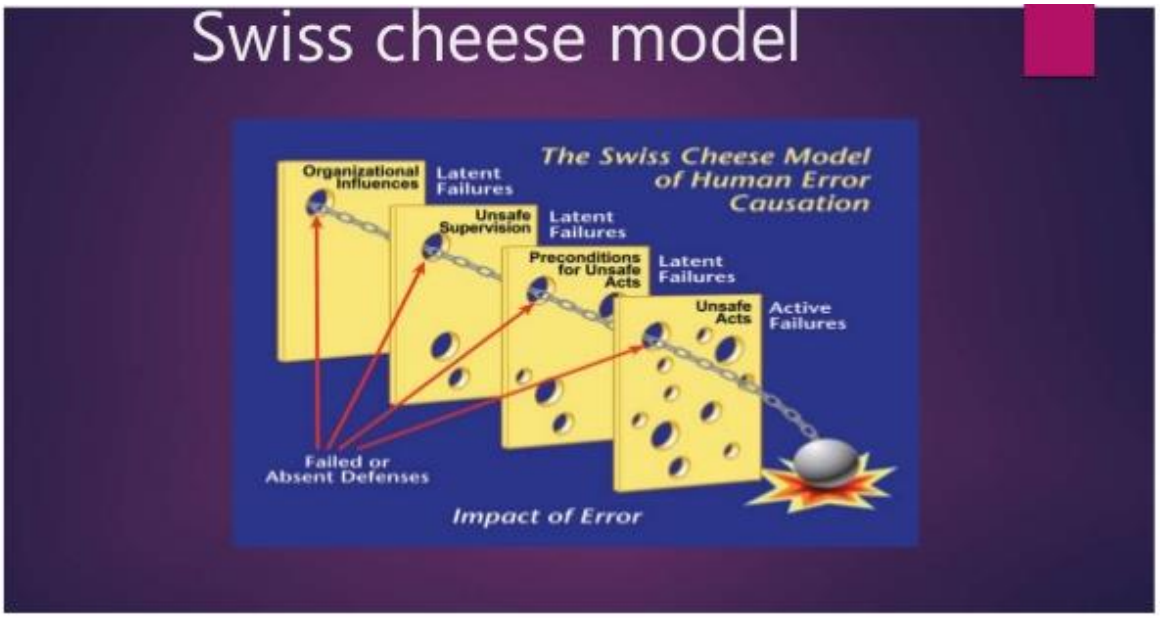


Figure 10: Reason's Swiss Cheese Model

The Swiss cheese model proposed by Reason in 1990 is, in fact, the linear chain-of-failure events model but drawn differently, using a different analogy (Swiss cheese versus boxes or dominos), and, in fact, less general. Like the Domino model (but unlike the general linear-chain-of-failure events model), it concentrates on human error, here called "unsafe acts" as in the Domino model. In this depiction of the chain-of-events model, the failure events are the holes in the cheese; an accident occurs if the holes or failure events line up. Note that there is still a chain of events and direct causality only. There are OR's in the Swiss cheese representation (multiple holes), but it is not clear how or whether an AND (combination of failure events) could be represented.

Whether the prevention measures or barriers are drawn next to the box as in Figure 8 or their failure is depicted as a box (labeled with the failure of the prevention device as in Bow Tie models) or the events or barriers are depicted as slices of cheese (as in Figure 10), the basic underlying causality model is identical: failure events (which includes failure of protection devices) occur in a linear (sequential)

order where the preceding event is necessary and sufficient for the following event to occur. Note that in the Swiss cheese depiction of this abstract model, the holes must line up and no hole can be skipped. The depiction of the chains of events as dominoes falling, holes in slices of Swiss cheese, or in other ways (such as men's bow ties) are only graphical differences. It is the same identical causality model.

It is not clear to me what Reason meant by a "latent failure." In engineering, failures cause erroneous states which in turn can lead to later failures. I have never seen engineers use the term "latent failure" to mean an erroneous state. Unless all the failures involved are coincidental in time (in which there is no chain but simply one box labeled with all the coincident failures), causal events always precede in time the effect of the event—usually called an error. So the term "latent" appears to imply simply a forward chain over time.

There are some differences between Figures 9 and 10 and the more general chain-of-events model depicted in Figures 7 and 8. The Domino and Swiss cheese depictions emphasize human error as the cause of events, particularly direct operational human error, over everything else. For example, if the wings fell off of a plane or a freak weather event occurs leading to a crash, it would appear, in the Domino Model, as somehow the result of a pilot's action along with the pilot's social ancestry or, in the Swiss Cheese Model, the pilot's actions would be assigned major responsibility although the pilot's actions might result from management "failures." Design errors do not seem to be included.

Hollnagel's Resonance Model and FRAM

Prof. Hollnagel has created what he describes as an "analysis method" called FRAM (Functional Resonance Analysis Method). He describes FRAM as a way to describe outcomes arising from the variability of everyday performance.[12] FRAM is not limited to accident outcomes, he claims, but can be used to "understand the behavior of any system." Let's examine some of these claims.

In his second book on Safety-II, Prof. Hollnagel says that FRAM

> *… is not exclusively a safety or risk analysis method but a method to analyse an activity in order to produce a model (description) of it. The FRAM can be used for safety analysis but also for task analysis, system design, etc.*

In fact, FRAM does not appear to be a "safety" analysis method or a risk analysis method at all or to be capable of performing these activities.

> *The purpose of the FRAM is to analyse how something is done, has been done or could be done in order to produce a representation of it. This representation is effectively a model of the activity because it captures the essential features of how something is done, using a well-defined format. In the case of the FRAM, the essential features are the functions that are necessary and sufficient to describe the activity and the way in which they are coupled or mutually dependent.*

This process is described in engineering as *specification* (using a specification language) and the result is a specification. Analysis methods in engineering analyze specifications, they do not create them. And "safety analysis," task analyses, etc. produce information about these properties from some type of system specification, they do not just involve creating that specification. I have not been able to find any analysis methods that can be used on FRAM specifications, although they might exist.

FRAM specifications use hexagons (6-sided shapes) to represent system functions, with each function described by time, inputs, outputs, control, preconditions, and resources (execution conditions). These are basically the same things that are specified in most every system specification created by decomposition of the functions, although the functions and properties may be depicted in different

[12] This explanation seems to contradict his claims, cited earlier, that *"as the world …becomes more complicated, it is no longer reasonable to assume that we can understand the causal relationships between actions and outcomes, or even that they can be described in causal terms."*

ways (boxes, circles, etc.) or simply written in words without a diagram. Note that this is classic decomposition, despite Prof. Hollnagel's claim (discussed earlier) that "*socio-technical systems are not decomposable*." The problem is that he starts with an incorrect definition of the basic properties of decomposition and emergence.

Figures 11a shows the FRAM specification of the steps in a simple cooking process. Figure 11b shows a small part of a FRAM specification of the steps in a healthcare process. They are both from Hollnagel's website. Note the complexity of the example in Figure 11b for what is a relatively simple healthcare process to identify patients.

In a FRAM specification, the hexagons (functions) are linked by the corners of one hexagon to the corners of other hexagons by lines with arrows to show the known and specified connections among the decomposed steps in the process. As will be seen in the section on Systems Theory, these are not, in fact, the only types of interactions or couplings that can occur between functions or components in a system. Much more subtle (and often unknown) relationships between functions can exist. As stated, a FRAM specification appears to involve standard functional decomposition.

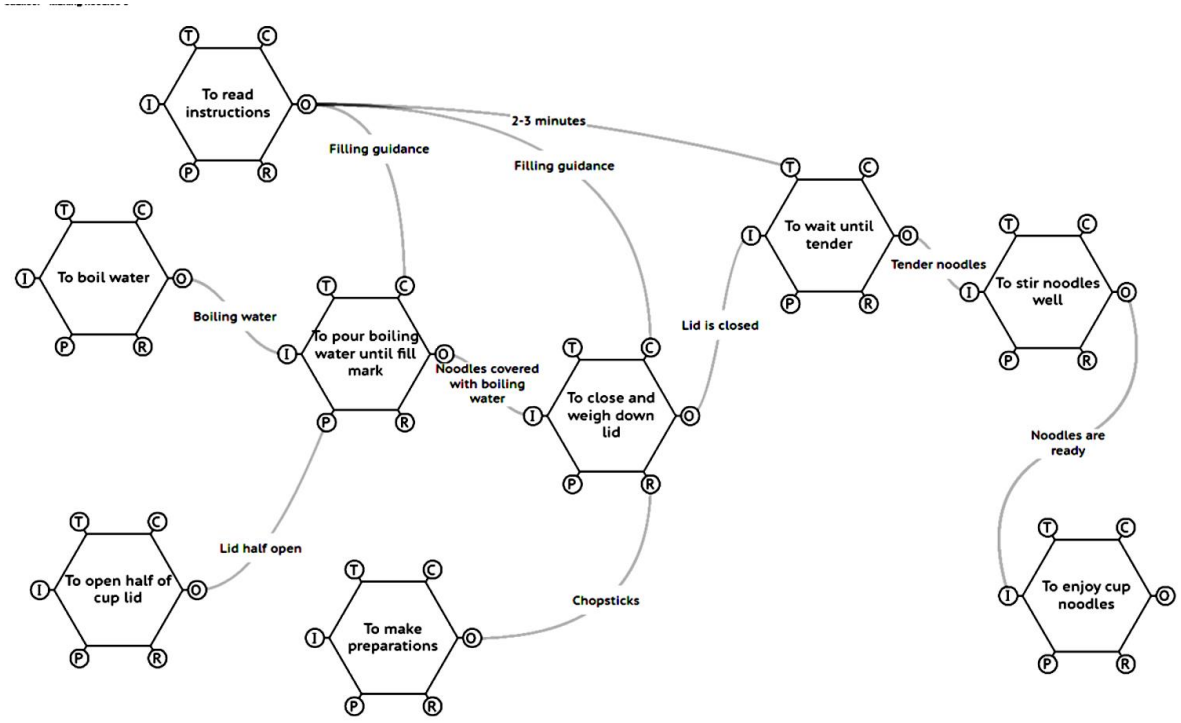


Figure 11a: An example of a FRAM specification of the steps in a cooking process. [Erik Hollnagel, FRAMSYNT: The Functional Resonance Analysis Method, https://functionalresonance.com/support/framsynt.html, p.14 and 22]

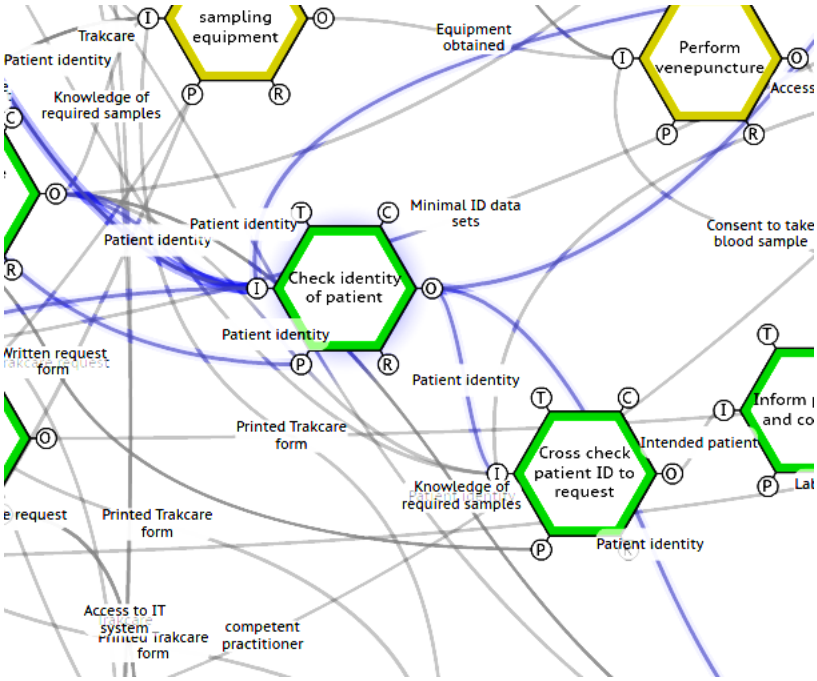


Figure 11:Two examples of a FRAM specification of the steps in a process. {Erik Hollnagel, FRAMSYNT: The Functional Resonance Analysis Method, https://functionalresonance.com/support/framsynt.html, p.14 and 22]

The FRAM specification language appears to be almost identical to SADT (Structured Analysis and Design Technique), which was created in the 1960s. Later, in 1981, SADT was renamed IDEF0 (Integrated DEFinition language). In SADT, the functions are written as rectangles rather the hexagons, and the lines are attached to the sides of the rectangles. FRAM has the two additional parameters of time and preconditions for the functions compared to the original SADT, but these two additional parameters, were later added to IDEF0 and are part of IDEF1 and IDEF2.

I could not imagine building a model of a very complex system, including all the technical and social factors, using FRAM or SADT. In fact, SADT and the various related IDEF specification methods have been used only for relatively simple business or manufacturing processes and for business (data processing and information system) software. They have not been found to be practical for complex real-time systems or products. The examples of FRAM that I have seen appear to be simply representations of the steps in relatively simple processes, as in Figures 11a and 11b. I cannot imagine how FRAM could be used to specify complex, safety-critical systems such as an aircraft, a nuclear power plant, a nuclear missile, or an automobile.

Beyond the problem of whether FRAM specifications are useful, particularly for safety, a more important question is what can be learned from them through analysis. Every analysis method related to safety must be based on some causality model or assumptions about how accidents occur. As there is no description of the underlying causality model on which FRAM is based, it is unclear what this causality model exactly is. In his second book, Prof. Hollnagel defines FRAM as a "*method sine model rather than a model cum method*" (a method without a model rather than a model together with a method). From a technical standpoint, this statement does not make much sense.

An analysis method is *always* based on some model or assumptions about system behavior. In engineering, an analysis method without an underlying model cannot exist, as Hollnagel claims for FRAM.

As an example, wing design is critical in maintaining lift in an aircraft, in other words, in keeping the aircraft up in the air. There are several analysis methods that can be used to calculate the lift in a particular wing design. These analysis methods are all based on underlying physical principles (a model), such as Bernoulli's principle, which says that if air speed increases then the pressure is lowered. Thus a wing generates lift because the air goes faster over the top of the wing, creating a region of low pressure, and thus lift. It would make no sense to have an analysis process or technique for evaluating the lift generated by a particular wing design that is not based on the underlying physics of flight.

For the humans in systems, there must also be a model used to explain behavior. For non-physical system components such as software, analysis methods are based on mathematical models, often based on formal logic or state-machine theory.

In hazard analysis, the underlying model includes assumptions about how accidents occur, whether that model is conscious or subconscious. In Prof. Hollnagel's case, this model may be that accidents are the result of a lack of resilience on the part of the human operators or perhaps the model involves functional resonance. Or maybe both; it is hard to tell when he denies that he is using a model but then says that things "go right" or "go wrong" using performance variability and his descriptions of hisconcept of "functional resonance." This is a model, just not a well-defined one.

The problem may arise from his confusing use of the term "model" to mean two different things. Prof. Hollnagel says that FRAM is a method to analyze an activity in order to produce a *model* of it. What he is describing is usually called the process of specification (not analysis) and the creation of a specification.

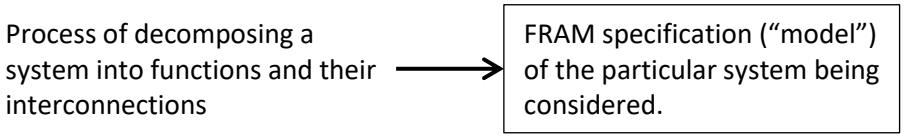


Figure 12: The FRAM process and "model"

In Figure 12, I have labeled the result of the decomposition/specification process as a FRAM specification/model, and I did not give a label for the process of creating the specification. Prof. Hollnagel uses the same name (FRAM) for both, which is very confusing and leads to fallacious arguments.

The term "model" when he uses it to talk about linear causality models is a much more general concept that applies to all systems, such as a linear causality model, not just the specification (or model) of a particular system. Analysis must always be based on some general model or assumptions about the thing being modeled (which is the definition of a model). The general engineering terminology and process is shown in Figure 13.

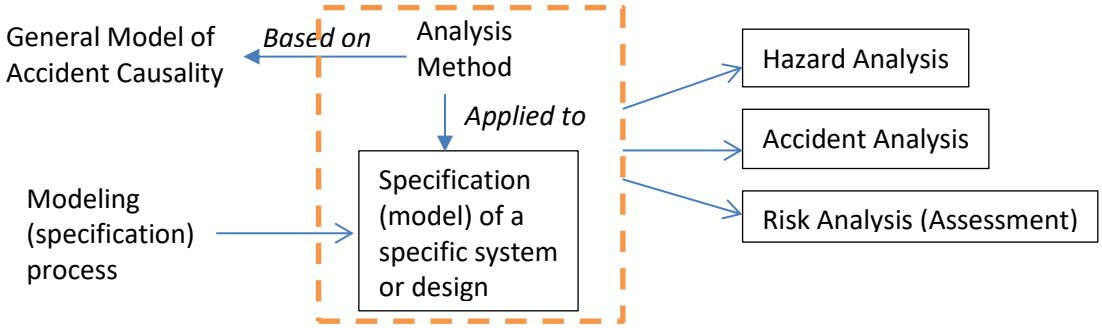


Figure 13: General process for creating safety-related analyses

Prof. Hollnagel claims in his second book that

The FRAM can be used for safety analysis but also for task analysis, system design, etc.

The process of creating a specification of something, shown in Figure 12, is very different than an analysis of properties of that thing. FRAM therefore is not a way to analyze a system to determine its properties, but it can be used for the creation of a specification of a particular system. So, examining the claim in the quotation above, how does safety analysis, task analysis, etc. take place using the FRAM specification? I have never seen any analysis methods that can be used to analyze a FRAM specification but perhaps they are documented somewhere.

Note that "safety analysis" here is undefined; instead engineers use the terms "hazard analysis" or "risk analysis," which have been carefully and formally defined. What would a "safety analysis" of a FRAM specification produce? The term "task analysis" in undefined here also, but if it means (as is common) simply writing down how the task is or should be performed, then that most likely can be easily produced from a FRAM specification.

I've thought about how FRAM specifications themselves could be analyzed and therefore whether a true analysis method *could* be created for such specifications. One way is to generate all the states of the system from the FRAM specification of the functions and their interactions and their defined interactions. While theoretically possible to generate all states, it is for all practical purposes impossible except for either very simple systems or specifications that omit all the important details about the system design. For example, we created a functional, black-box specification of TCAS (a collision avoidance system for aircraft) for the FAA during its certification. From the specification, we determined that the system had at least 10^{40} states. Because our specification had a mathematical foundation, analysis could be performed on it, but I don't see a mathematical foundation underneath FRAM—simply a diagramming technique. Alternatively, it might be possible to simulate a system specified by FRAM—in fact, IDEF specifications can be simulated—but only a very small number of the potential states of the system could be examined practically in this way.

Adding to the confusion is that while declaring that FRAM is not based on a model, he says that FRAM "*does not refer to an existing model but rather to four principles or assumptions about how things happen*." In science and engineering, this is the definition of a model. So let's examine them with the assumption that this is the model of behavior underlying the FRAM specification process (or underlying FRAM specifications[13]).

The four principles are

1. The equivalence of successes and failure
2. Approximate adjustments
3. Emergent outcomes
4. Functional resonance

Let's look at each of these.

> ***First principle: the equivalence of successes and failures*** *Explanations of accidents and incidents typically rely on decomposing a system or an event into parts, either physical parts such as people and machines or the segments of an activity such as individual actions or steps in a process. Outcomes are explained by linear cause-effect relations among the parts, and adverse outcomes are attributed to malfunctions or failures of parts.*

[13] I am purposely using the term "specification" for FRAM diagrams to avoid the same problem that is in Prof. Hollnagel's books: an overloading of the term "model" and moving from one definition to the other for convenience in making a particular argument.

Linear cause-effect relations are among *events*, not among "parts" of systems. Decomposing systems into parts or steps in a process is exactly what FRAM does, not events. As Hollnagel states, attributing outcomes to malfunctions or failures of parts is one way to define linear cause-effect relations where the cause is a failure or malfunction, as in the linear chain-of-failure events model, the Domino model, and the Swiss Cheese model. It is also the basis for causal analysis techniques based on these models such as fault tree analysis, event tree analysis, failure modes and effects criticality analysis, etc. But it is not the only way to define cause-effect relations in linear models.

As an example, HAZOP uses an underlying model of causation that says the effects arise from deviations from the design or operating intentions, which is more general than just failures. But it is still linear. Prof. Hollnagel appears to be arguing in his books that the effects can created as a result of performance variability (which sounds very similar to the HAZOP definition). Defining the effects as the result of performance variability (vs. failures) does not change the fact that the relationship between cause and effects is linear. He writes:

> *This implies that the causes of things that go wrong are different from the causes of things that go right. Otherwise the endeavour to 'find and fix' the causes of unacceptable outcomes would also affect the occurrence of acceptable outcomes.*

The illogic here was examined earlier in this paper.

> *The FRAM – and Resilience Engineering – takes a different approach, namely things that go right and things that go wrong happen in much the same way.*

This seems to be saying simply that all causes—no matter whether the results are desired or undesired—are the result of performance variability. This could be a possible definition of causality, but some examples would have been nice because I cannot think of many instances in which this assumption, that it explains both things that "go right" as well as things that "go wrong" would hold. And I can think of lots of counterexamples. For example my brakes may not stop the car because their performance varied from the design. In fact, this is the definition of going wrong in almost all cases. But, at the same time, the reason my brakes *did* stop my car is not because of performance variability—the behavior resulted because the performance did *not* vary from the design or at least the requirements. At best, the statement applies only to human behavior, but even there the conclusion is suspect or perhaps trivial. It's hard to tell with all the vague and undefined terminology being used here and no examples.

To me, if things happen in exactly the same way, then the results must be the same unless we live in a completely random world. More practically, I do not understand how successes and failures can be defined as having the same cause (happening in the same way) in any useful accident causality model.

> *The fact that the results are different does not mean that the explanations must be so as well. The principle of approximate adjustments explains why this is so.*

If the results are different, then the explanations are not the same. Explanations are used to explain why the results occurred. "Approximate adjustment" describes a type of behavior; it is not an explanation of causality.

In a specific instance, if the results are different, then the explanations must be so as well. For example, the explanation that a particular human behavior varied outside the boundaries of safe behavior (that is, a hazard occurred while the human was adjusting performance) is not the same as an explanation that says the human behavior did *not* go outside the limits of safe behavior while adjusting performance (no hazard occurred while the human was adjusting performance). The results are different so the explanation for the results in the particular case must be different. Otherwise, the results cannot be different. If the operator's behavior did not vary in any way, then it must not have any impact on the accident and the causes must be found elsewhere.

I assume, as I speculated earlier in this paper, that Prof. Hollnagel is simply saying that all explanations of how things "go right" and how things "go wrong" involve approximate adjustment, i.e., all behavior can be explained using this concept. This fact, even if we accept it, does not mean that successes and failures are equivalent or the causes of "going right" and "going wrong" are equivalent but only that he is using a causality model that assumes all causes involve performance variability and, more specifically, that variability is based on approximate adjustments. There is a difference between a general assumption about causality and the causes in particular losses.

> ***Second principle: approximate adjustments*** *Sociotechnical systems cannot be specified in minute detail because humans are not machines. Effective work requires that performance continuously is adjusted to the existing conditions (resources, time, tools, information, requirements, opportunities, conflicts, interruptions). Adjustments are made by individuals, by groups and by organisations and take place at all levels, from the performance of a specific task to planning and management. Since resources (time, materials, information, etc.) are almost always limited, the adjustments will typically be approximate rather than precise. This is rarely critical because people will know what to expect and be able to compensate for that. The approximate adjustments are the reason why things mostly go right and why they occasionally go wrong.*

The first sentence, "*Sociotechnical systems cannot be specified in minute detail because humans are not machines*," again assumes that sociotechnical systems are only comprised of humans and the "socio" parts. The technical parts and even the social parts have components that are not humans. So this assumption (and the resonance model) applies only to systems that include nothing but humans. There are few if any of these today, particularly safety-critical ones.

If we assume that we can explain and prevent all accidents by looking only at human controllers, the rest of the description of this principle makes sense to me until the end. Of course humans adjust their behavior to the existing conditions. Who would believe otherwise? However,

> *"…the adjustments will be approximate rather than precise. This is rarely critical because people will know what to expect and be able to compensate for that.*

The last two sentences are problematic. That people "*will know what to expect*," particularly when it has never happened to them before, is doubtful. The second part of that sentence, i.e., that they will "*be able to compensate for that*" is incredibly optimistic. Even if they know what has gone wrong, the ability to compensate usually depends on the design of the system that they are controlling. If all the engines fail on a plane, even if the pilot knows what has happened, he or she will probably not be able to compensate for it. While an explanation using performance variation and "functional resonance" might apply for a few accidents, it is very far from a general accident theory. It is puzzling why no examples of how his functional resonance model explains real accidents that have occurred are included in either of these two books on Safety-II.

> ***Third principle: emergent outcomes*** *The variability of individual functions is rarely large enough to serve as the cause of something going wrong or to be described as a failure. The variability of multiple functions may on the other hand combine in unpredictable (nonlinear) ways that can lead to unexpected and disproportionate outcomes – negative as well as positive. Acceptable and unacceptable outcomes can both be explained as emerging from variability due to the everyday adjustments rather than as a result of single or multiple cause-effect chains starting from a malfunction or failure of a specific component or part.*

As discussed above, this is not the standard definition of emergence used in science and system theory. Emergent outcomes are usually predictable and expected. Non-linear is not equivalent to unpredictable in either the mathematical definition of "non-linear" (as used in the above paragraph) nor the definition of the term as meaning non-sequential as used in accident models. Causality may be non-sequential but

still anticipated and predicted. In fact, we design systems that depend on emergent properties. For example, we design air traffic control systems to optimize throughput, which is an emergent property. The throughput can be calculated and optimized. Finally, "variability due to everyday adjustments" simply sounds like feedback control, one of the basic concepts in engineering.

> ***Fourth principle: functional resonance*** *As an alternative to linear causality, the FRAM proposes that the variability of two or more functions can coincide and either dampen or amplify each other to produce an outcome or output variability, that is disproportionally large.*

One problem with this claim is that FRAM is not an alternative to linear causality. His principle of resonance still assumes linear (or sequential) causality in the non-mathematical (sequential) sense. The different outcomes are the events. He is confusing the two uses of the term "nonlinear." Consider the linear causality chain in Figures 7 and 8 where moisture leads to corrosion, which leads to weakened metal. The *mechanism* for the causation is not specified but is hidden in the arrow. That is where resonance/approximate adjustments would lie, in the arrow but not in the events caused by the resonance. The difference appears to be in the way that the linear/sequential chain progresses, in this case, by supposedly increasing variability. But there is still a linear/sequential chain of events leading to a loss. Perhaps Prof. Hollnagel is confusing events with the system components or parts as noted earlier?

The more important problem, however, is that functional resonance is not a general model of causality in complex systems unless it can explain all accidents and I do not see how that is possible. Prof. Hollnagel conveniently does not provide any examples, but it is not a reasonable explanation for the sequence of events in the accident shown in Figures 7 and 8. For example, corrosion causing weakened metal in the tank is not the result of approximate adjustments. If I am daydreaming while driving and do not notice that the car in front of me has stopped and do not activate the brakes, how is that approximate adjustments? Or I am not daydreaming and I do notice the car in front as stopped and I do activate the brakes, how is that "success" a result of approximate adjustments or resonance?

Perhaps the closest to Prof. Hollnagel's Resonance Model, as noted earlier, is the model of accident causation underlying the HAZOP analysis method. This model still involves a linear chain of events, but instead of the events only being caused by general "failures" of components, they are caused by "deviations" in the operating parameters. Of course, the difference depends on the definition of "failure" being used. In HAZOP, the parameter deviations lead to failures. No matter what the mechanisms in generating the outcome(s) from the preceding event, e,.g., component failures, parameter deviations, or approximate adjustments, the sequence of outcomes or effects is linear.

> *In the latter case, the consequences may spread to affect other functions in analogy with the phenomenon of resonance. Functional resonance describes the noticeable performance variability in a sociotechnical system that can happen when multiple approximate adjustments coincide. Performance variability is not random because the approximate adjustments comprise a small set of recognizable adjustments or heuristics. There is some regularity in how people behave and in how they respond to unexpected situations – including those that arise from how other people behave. Functional resonance offers a systematic way to understand outcomes that are both noncausal (emergent) and nonlinear (disproportionate).*

It is hard to understand how this can be "noncausal." In fact, the cause is stated explicitly here—the cause is performance variability, which in turn is hypothesized to be caused by approximate adjustments. Emergent outcomes all have causes.

"Nonlinear" (disproportionate) outcomes, as used here, sounds like he is thinking of the mathematical definition of "non-linear" and not the definition used in linear causality models, which has to do with the sequentiality of the events in the outcomes.

Engineers and scientists create formal definitions for concepts and terminology. The reason is to avoid the serious problems inherent in the arguments in these two books on Safety-II. Neither book includes a formal definition of functional resonance (or most of the rest of the terminology). Perhaps it is elsewhere?

Prof. Hollnagel says that he derived his concept of functional resonance from the concept of resonance in Physics. The Physics concept is formally defined. In Physics, resonance is the tendency of a system to vibrate with increasing amplitudes at some frequencies of excitation. The principle of resonance in physics states that when two frequencies are brought together, the lower will always rise to meet the higher. More recently, the concept of stochastic resonance has been described where the output is not proportional to the input. In mathematics, this is the definition of "non-linear."

Non-linearity is important because most systems are inherently nonlinear in nature. However, the term linear (and therefore non-linear) is used very differently in causality models, where, as described above, it simply means that causality is described as a chain (sequence) of events where each event is defined by a necessary and sufficient" relationship with a preceding event. "Sequential" would be a better term to use but it is hard to change terminology after it has become established.

Note that Prof. Hollnagel (and other social scientists) frequently use the terms "blunt end" and "sharp end, but this implies an underlying assumption of linearity: There are only "ends" if we have a line. Circles, for example, do not have ends. Neither do events occurring simultaneously. There is an unrealistic simplicity in the use of the terms "blunt end" and "sharp end": for example, the decisions made by the aircraft designers have as much impact on whether the aircraft stays aloft as the pilot's decisions (perhaps more).

Prof. Hollnagel claims that his concept of functional resonance is based on stochastic resonance but is different in that the changing behavior is based on approximate adjustments. The relationship to the engineering/physics use of the term resonance is not really clear. He writes:

> *Functional resonance differs from stochastic resonance because the emergent effect is based on approximate adjustments rather than random noise. The variability of performance in a socio-technical system represents the approximate adjustments of people, individually and collectively, and of organizations that together constitute everyday functioning. Because these approximate adjustments are purposive and rely on a small number of recognisable short cuts or heuristics, there is a surprisingly strong regularity in how people behave and in the way they respond to unexpected situations—including those that arise from how other people behave.*

The analogy with physical resonance seems misplaced here. Of course, people are not random in their behavior. The whole field of psychology is based on this assumption. But it is overly simplistic to describe all human behavior as approximate adjustments (and rather surprising for a psychologist to do so). People do try different solutions when they come across a problem, and they also try to improve their performance over time. But to describe all human behavior in this way or to define approximate adjustments based on heuristics as the cause of all accidents is an extreme oversimplification. Humans always have multiple goals and they can be conflicting (and even subconscious).

I could find no examples of specific accidents using functional resonance in either of his books. It would be helpful to see how this model and concept could explain a large number of accidents or even one—although I am sure I could find one that it explains as it seems to be simply the concept of feedback control in engineering. Perhaps it could be used to explain accidents involving pilot-induced oscillation (PIO)?

The biggest problem, of course, is that only humans and organizations are considered in his book, which seems to imply that accidents have nothing to do with the software and hardware in the system, i.e., this is not a sociotechnical approach but simply a "socio" one. The technical is ignored in both books. Engineers do not make most design decisions using short cuts or heuristics (as functional

resonance is explained above). Physical devices such as brakes and engines certainly do not work by successive approximate adjustments. How does this model explain an accident such as the brakes in your car failing, perhaps because of a brake fluid leak, and the car hitting an object ahead? Where are the approximate adjustments?

It would have been very useful if at least one complex accident had been explained in the book using this model, particularly one not involving only human operators. The only way I can think of using this model to explain accidents would result in placing all "blame" on the human operators and perhaps on management and almost surely oversimplifying the cause.

Prof. Hollnagel says:

> Page 150: *"explanations that are based on linear causality are, however, dangerously oversimplified. Because the events are rare and because they are difficult to explain and understand*, *they are also difficult to change and manage.*

First, I agree with the statement that linear causality dangerously oversimplifies the explanation of accidents, but that includes his resonance model. Not only is linear causality not powerful enough to include many of the most important factors in accident, some of the linear causality models like the Swiss Cheese, Domino, and Prof. Hollnagel's resonance model focus only on humans as the cause of accidents.

I disagree, however, with the statement that events are difficult to change and manage because they are rare and because they are difficult to explain and understand. Why? The events in most accidents are not difficult to explain and understand. They are rare because systems are designed to prevent them. This is exactly what all accident prevention methods are about, i.e., preventing the events that lead to hazardous states and accidents. Accidents are rare because these efforts are usually successful, not necessarily because they are difficult to explain and understand. And I'm not sure what any of this has to do with explanations of linear causality being oversimplified.

In summary, I have no experience with FRAM nor do I know anyone who has used it on a complex system, so I cannot evaluate it. I cannot find an example for a complex real system that has been published or is on the internet, although I have looked. It sounds, however, like it involves analytic decomposition of the system functions, where an assumption is made that couplings and interdependencies are limited and understood and therefore can be specified. Accidents, in fact, often occur when these interactions are not understood using linear causality assumptions and need more powerful accident causality models underlying the analysis methods.

If it really were feasible to completely model the behavior of complex systems (both all the things that "go right" and all the things that can "go wrong"), wouldn't that solve all our problems in engineering? Wouldn't we have done it long ago? It couldn't be done using SADT/IDEF, so why would it be possible with a specification language that is virtually identical?

Limitations of the Linear Chain of Events Model in General

Most of safety engineering has been built on the linear chain-of-events causal model of how and why accidents occur, although not the more limited Domino, Swiss cheese, and functional resonance models that focus only on human error.

The example accident in Figures 7 and 8 is quite simple. Real systems today may have hundreds and probably thousands of such chains of events leading to losses. I was told of one fault tree analysis (which generates linear causal chains, just as do all traditional hazard analysis techniques) for one aircraft's Integrated Modular Avionics system that required over 2000 pages to document the results. And this was only for one part of the aircraft. During Space Shuttle development, a FMEA (Failure Modes and Effects Analysis) identified 40,000 critical items. It's not clear what to do with the information that the failure of 40,000 individual items could lead to a serious loss, but only a government project like the

Space Shuttle could have the resources to identify all of these, let alone provide protection against them. And, of course, Space Shuttle design errors and poor management decision making are omitted from this analysis: this omission includes the causes attributed to the two actual Space Shuttle losses.

Abstractions or models are used by humans to understand complex phenomena. By definition, they leave out factors (otherwise they would be the thing itself and not useful). For abstractions or models to be useful, they need to include the important factors or factors of interest in understanding the phenomenon and leave out the unimportant. Unfortunately, the simple linear chain-of-events causality model leaves out too much to be useful in understanding and preventing accidents in complex sociotechnical systems.

There are other inherent limitations of this traditional and almost universal chain-of-failure-events model. First, there is an assumption that the events and barriers fail *independently*, that is, there is nothing that will reduce or eliminate the effectiveness of all of them at the same time. In the Swiss Cheese Model (where the barriers are depicted as Swiss Cheese Slices), the slices are assumed to be independent. Given this assumption, the risk of an accident, if all defenses are implemented correctly, is theoretically low. However, the independence assumption is almost always untrue in real systems. For example, accidents commonly occur because budget cuts, demands for increased productivity, or competitive pressures make all the "barriers" (protections) ineffective at the same time. A poor organizational or safety culture (e.g., management pressures to ignore safety rules and procedures) can also undermine the effectiveness of all the safety controls and the applicability of the model. These so-called "systemic factors" do not appear in linear-chain-of-event models and, for the most part, such factors have to be ignored to perform quantitative or even qualitative risk assessment based on linear causality.

Another critical omission from the linear chain-of-failure events model are accidents that involve non-failures, where all the components may operate as designed, but their interactions lead to a "failure" of the system as a whole. Accidents resulting from the unsafe interaction of non-failed components, i.e. that operate according to their specification, may stem from complexity in the overall system design and the incorporation of software controls and autonomy in the design. What system components failed in the Warsaw A320 reverse thruster accident described earlier? Certainly not the flight crew or the software, both of which did exactly what they were instructed to do.

Most of the accidents I see today are a result of these types of system design errors, although they are often incorrectly blamed on the pilots or human operators in general. That does not mean, as Prof. Hollnagel claims, that these accidents cannot be explained or that there is no cause involved. After the fact, they can be explained. The problem lies in trying to predict them using the traditional hazard analysis techniques that assume accidents are caused by chains of component failures. More powerful causality models and hazard analysis techniques built on them are needed to identify such accident causes in a hazard analysis.

Consider the release of methyl isocyanate (MIC) in Bhopal India in 1984, which is considered the worst industrial accident of all time. Tens of thousands of people were killed and even more suffered debilitating injuries. There was no lack of protection devices and barriers to prevent this accident. Because of the dangers associated with the production of MIC, primarily related to contact of the chemical with water, many physical barriers are used to prevent such contact, including slip discs in valves, concrete around storage tanks, and operational procedures. Those are the first level of barriers in the defense-in-depth approach taken. In addition, if the water did get through the barriers, which would cause an enormous increase in pressure and heat, there were relief valves, procedures specifying that the tanks were never to be more than half their maximum capacity along with spare tanks to bleed the contents of one tank into an empty one, a refrigeration unit to limit the reactivity of the MIC and a high temperature alarm along with lots of other gauges and instrumentation to keep the operators informed. If, despite all these protections, a release of MIC occurred, there was a vent scrubber to

neutralize any escaping gas with caustic soda, a flare tower to burn off any escaping gas missed by the scrubber, and a water curtain to knock down any gas missed by both the scrubber and the flare tower. There was, of course, also warning sirens, protective equipment, and frequent testing of the alarms and practice of emergency procedures.

Despite all these preparations and a defense-in-depth design, tens of thousands of people were killed in an inadvertent release of MIC. How could the vent scrubber, flare tower, water spouts, refrigeration unit, alarms and monitoring instruments, etc., all fail simultaneously? In fact, a probabilistic risk assessment of this plant would have combined the probability of the "failure" of all these devices and come up with an extremely low (basically impossible) likelihood.

The answer is that a defense-in-depth strategy does not protect against systemic factors that impact all the barriers and protection devices (Swiss cheese slices or boxes in the event chain) at once, as described earlier. At Bhopal these systemic causal factors included design errors in most of the protection devices and severe pressures on the company to cut costs due to a sharp decrease in demand for MIC. The cost cutting pressures led to cutting maintenance and operating personnel in half and reducing maintenance procedures (the scrubber and flare tower were out of operation at the time of the accident because of a lack of maintenance), turning off the refrigeration unit to save money, unskilled workers replacing skilled ones, minimal training of workers in how to handle emergencies, and reductions in educational standards and staffing levels, among other things.

There had even been warnings in the form of a less serious accident the year before, several serious incidents involving MIC in the previous three years, and an audit report two years before that noted all the deficiencies in the plant that led to the loss. The deficiencies in the audit report were never corrected. The more one learns about this accident—and, indeed, most accidents—the less important the role of the operators appears to be in causing or not preventing the accident. In this case, no matter how adaptable or resilient the operators had been, they could not have prevented the tragedy. Using too simplistic an accident causality model can lead to tragedies.

Past assumptions about the role of humans in systems also do not fit systems today, where the humans are mostly managing complex automation rather than directly controlling physical devices or computer-automated functions. The future will see even more changes to the human role away from active control and toward being a manager or monitor of computers and even partnering with automation to achieve common goals as responsibilities are divided between the machine and the human. Autonomy does not usually mean that humans are totally eliminated from systems (except in the simplest cases) but only that their roles are changed. None of these new roles and human factors considerations are included in the traditional model of accidents caused by chains of failure events. They cannot be represented using a simple linear failure model—including one that explains accidents as performance variation/approximate adjustments.

Again, note that descriptions of chains of events as dominoes falling, holes in Swiss cheese, similarities to men's formal attire such as bow ties, or approximate adjustments are only graphical differences. The chains of events may be drawn differently, use different notations, or apply different analogies, but they all are describing the same underlying chain-of-events model. They are not different causal models, but simply different names or notations for the same model.

Safety engineering has been built on this limited causal model of how and why accidents occur. In a FRAM specification, the hexagons (functions) are linked by the corners of one hexagon to the corners of other hexagons by lines with arrows to show the known and specified connections among the decomposed steps in the process. As will be seen in the section on Systems Theory, these are not, in fact, the only types of interactions or couplings that can occur between functions or components in a system. Much more subtle (and often unknown) relationships between functions can exist.

Despite these limitations, no alternative to this traditional linear accident causality model has been suggested until relatively recently except for Epidemiological models.

Epidemiological Models

Prof. Hollnagel labels the Swiss cheese model as an epidemiological model, but the Swiss cheese model has little relationship to the basic science underlying the field of epidemiology. Historically, there has been a true epidemiological model of accident causality. In the 1940s, John Gordon, an professor of epidemiology at Harvard University, stressed the multifactorial nature of accidents [Gordon, 1948]. He and others suggested that accidents should be viewed as a public health problem that can be handled using an epidemiological approach. In the epidemiological model of causality, accidents are conceptualized in terms of an agent (physical energy), the environment, and the host (victim). Accidents are considered to result from complex and random interactions between these three things and cannot be explained by considering only one of the three factors or by simple linear interactions between events.

Two types of epidemiology have been applied [Thygerson, 1977]:

- Descriptive epidemiology: The general distribution of injuries in the population is described by determining the incidence, prevalence, and mortality rates for accidents in large population groups according to characteristics such as age, sex, and geographical area.
- Investigative epidemiology: The specific data on the causes of injuries is collected in order to devise feasible countermeasures.

The epidemiological approach to accident causation assumes that some common factors are present in accidents and that these can be determined by statistical evaluation of accident data, in much the same way that epidemiology handles disease. Because specific relationships between factors are not assumed, previously unrecognized relationships can be discovered. A claim is made that determinant as opposed to chance relationships can be distinguished [Benner 1984; Manuele, 1984].

This model has not been widely used. The validity of the conclusions from such epidemiological studies of accidents is dependent on the quality of the database used and the statistical significance of the anomalies found in the sample [Hope, 1983]. In practice, the data reported by accident investigators may be limited or filtered. Also, the sequencing and timing relationships between events and conditions is not captured by a purely statistical approach. This might be considered a limitation or a feature, but sequencing and timing relationships can provide important information when considering causality.

Systems Theory and STAMP

What is the alternative to linear/sequential causality chain models? One alternative is STAMP, which expands the linear model to more complex causes of accidents [Leveson, 2012]. Our models need to be able to explain all or at least most accidents in current systems to be useful. STAMP is the causality model that underlies what is called Safety-III in this paper.

Prof. Hollnagel writes the following, which is completely wrong:

> *p. 106: The Systems-Theoretic Accident Model and Processes (STAMP) approach has been developed by Nancy Leveson … It is a composite linear analysis method based on a view of systems as being hierarchically structures. The properties that emerge from a set of constraints at one level of hierarchy are controlled by constraining the degrees of freedom of those components, thereby limiting system behavior to the safe changes and adaptations implied by the constraints*.
>
> *p. 96 Other, more complicated but still linear schemas or explanation can be found in the guise of specific methods or approaches, such as Tripod, AcciMap, or STAMP. … And STAMP uses a model of sociotechnical control comprising two hierarchical control structure as a basis for its explanation.*

Prof. Hollnagel is incorrect here. STAMP is *not* an analysis method, it is an accident causality *model*. It is *not* linear—it does not model losses as chains of failure events. In addition I cannot understand where

Prof. Hollnagel got the impression that the model uses two hierarchical control structures, but perhaps that confusion is the result of seeing only one example that happened to include two control structures (Figure 17 included below in this paper). In fact, STAMP is a theoretical causality model, not a method like Tripod and Accimap. There are several general analysis methods built on STAMP, for example, STPA is used for general hazard analysis and CAST is used for retrospective accident analysis.

A Brief Introduction to Systems Theory

STAMP is based on the mathematical/theoretical foundation of Systems Theory. System Theory was created after World War II to deal with the new technology and complexity in our engineered systems. It arose at the same time in biology, where the complex interdependencies of the human body also created a need for a paradigm change to make significant progress. Before this time, complexity had been handled in science and engineering for several hundred years by using analytic decomposition, i.e., breaking systems into their components, analyzing the components independently, and then combining the results to evaluate and understand the behavior of the system as a whole (Figure 14). Events were treated as chains of cause-effect relationships.

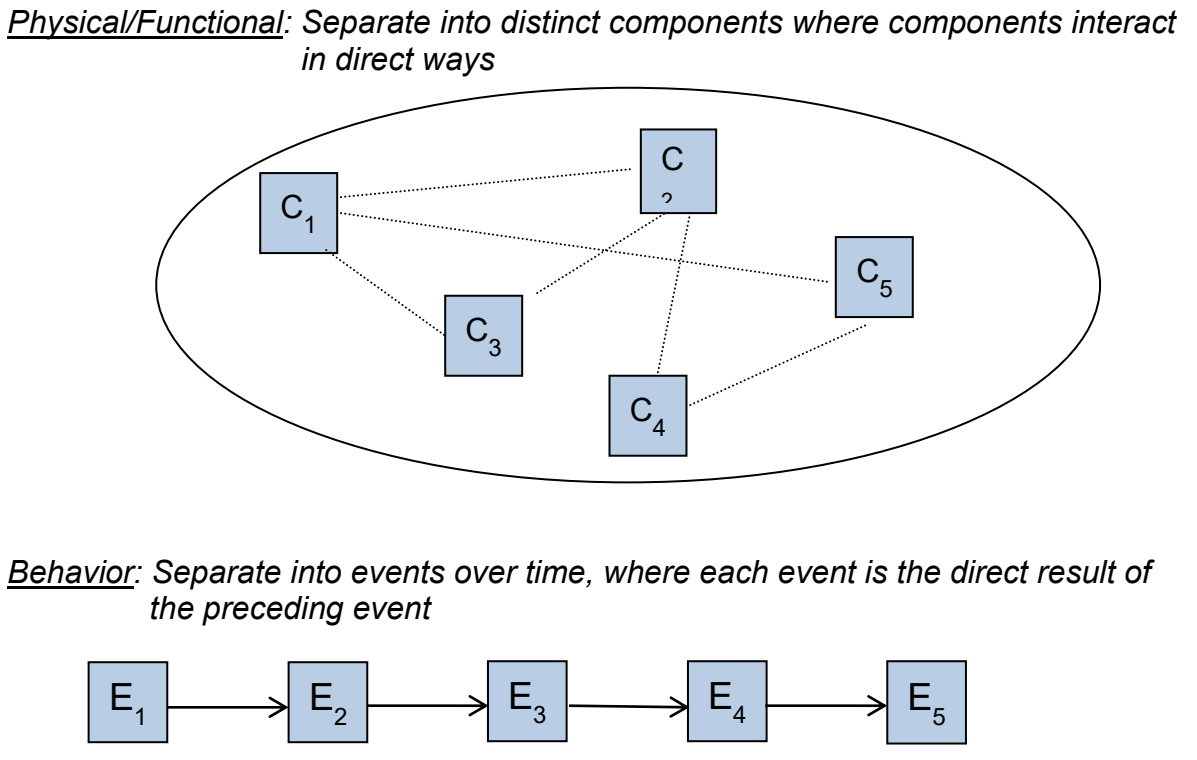


Figure 14: Analytic decomposition, shown at the top for system components and at the bottom for behavior over time.

In analytical decomposition, the physical or functional components are assumed to interact in direct and known ways. For example, if the weight of an complex object is the analysis goal, the separate pieces may be weighed and the result combined to get the system weight. As another common example, the system reliability is usually calculated by evaluating the reliability of the individual components and then the component reliabilities are combined mathematically to calculate the system

reliability. Safety is usually assumed to be simply a combination of the component reliabilities. For analytical reduction to work, the interactions and interfaces between components are assumed to be identifiable, as is true for FRAM models.

The limitations of analytic decomposition were described earlier in the section on Decomposition and Emergence.

Systems Theory as used in engineering was created after World War II to deal with the increased complexity of the engineered systems being built [Weiner,1965; Checkland,1981, Weinberg, 1975] and to understand the complexity of biological systems [von Bertalanffy, 1969]. In these systems, separation and analysis of separate, interacting components (subsystems) distorts the results for the system as a whole because the component behaviors are coupled in non-obvious ways. The first engineering uses of these new ideas were in the missile and early warning systems of the 1950s and 1960s.

Some unique aspects of System Theory are:

- The system is treated as a whole, not as the sum of its parts. You have probably heard the common statement: "The whole is more than the sum of its parts."
- A primary concern is *emergent properties*, which are properties that are not in the combination of the individual components but "emerge" when the components interact (Figure 15). Emergent properties can only be treated adequately by taking into account all their technical *and social* aspects. Safety and security and most other important system properties are emergent.
- Emergent properties arise from relationships among the parts of the system, that is, by how they interact and fit together.

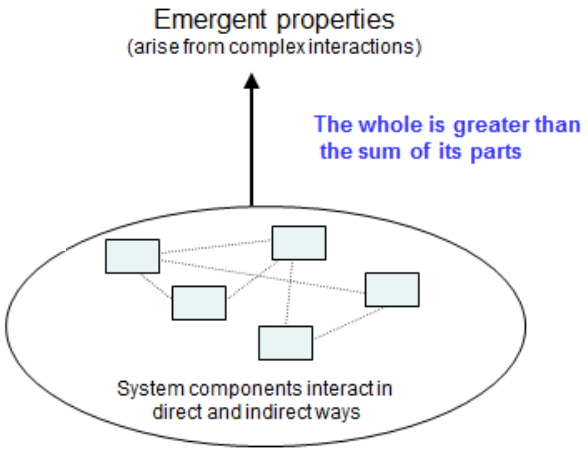


Figure 15: Emergent properties arise from complex interactions

If emergent properties arise from individual component behavior and from the interactions among components, then it makes sense that controlling emergent properties, such as safety, security, maintainability, quality, and operability, requires controlling both the behavior of the individual components and the interactions among the components. We can add a controller as shown in Figure 16.

The controller provides control actions on the system and gets feedback to determine the impact of the control actions. In engineering, this is a standard feedback control loop. In other words, control actions are issued by the controller on the controlled process. Feedback is used to determine the effect of previous control actions and whether additional control actions are required. In this way, the system

is treated as an adaptive system that is kept in a state of equilibrium over time as the system and environment change.

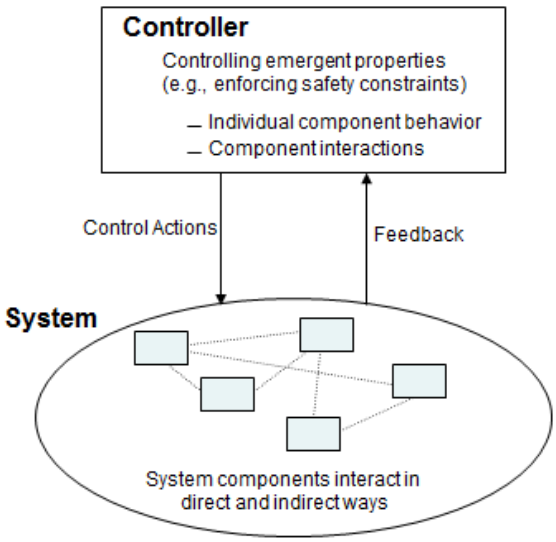


Figure 16: Control of emergent properties

The simplest example that most everyone is familiar with is a thermostat. The thermostat is provided with a setpoint and gets feedback about the current temperature in the controlled space. If the temperature is below the setpoint, then heat is applied until feedback (sensors) show that the setpoint has been reached. Then the application of heat is discontinued. The opposite is done if the measured current temperature is above the setpoint. Note that variability and adaptability are not precluded in this concept; indeed, they are assumed. For the thermostat, the temperature is assumed to vary over time (due to a variety of influences in the system and in the environment). The thermostat is used to limit the variability to an acceptable range or to adapt when a different setpoint is entered. An acceptable range, in the case of safety, is one that does not include hazardous states.

A more complex example is throughput in an air traffic control system. If each airline tries to optimize their routes independently, chaos would result as more aircraft might arrive at a popular hub airport than there are runways and gates available. In each airline's attempt to optimize their routing, most would find that they were unable to achieve their optimum and overall throughput in the system would suffer. By introducing an air traffic control system, each airline might not be able to get their optimum schedule, but the system throughput is optimized and each will hopefully do better than they would have otherwise. Schedules and traffic are optimized for the U.S. National Airspace using a central computer.

The controller enforces constraints on the behavior of the system, both safety and other types of constraints. Example safety constraints might be that aircraft or automobiles must remain a minimum distance apart, pressure in deep water wells must be kept below a safe level, aircraft must maintain sufficient lift to remain airborne unless landing, accidental detonation of weapons must be prevented, and toxic substances must never be released from a plant. These are standard types of hazards that we try to prevent in designing and operating systems.

Control is interpreted broadly and, therefore, includes everything that is currently done in safety engineering plus more. For example, component failures and unsafe interactions may be controlled through design, such as using redundancy, interlocks, barriers, safety margins, fail-safe design, etc. Safety may also be controlled through process, such as development processes, manufacturing

processes and procedures, maintenance processes, and general system operating processes. Finally, safety may be controlled using social controls including government regulation, culture, insurance, law and the courts, or individual self-interest. Human behavior can be partially controlled through the design of the societal or organizational incentive structure and not simply by issuing rules and procedures.

To model complex sociotechnical systems requires a modeling and analysis technique that includes both social and technical aspects of the problem and allows a combined analysis of both. Figure 17 shows an example of a hierarchical safety control structure for a typical regulated industry in the U.S. (international controllers could have been included).

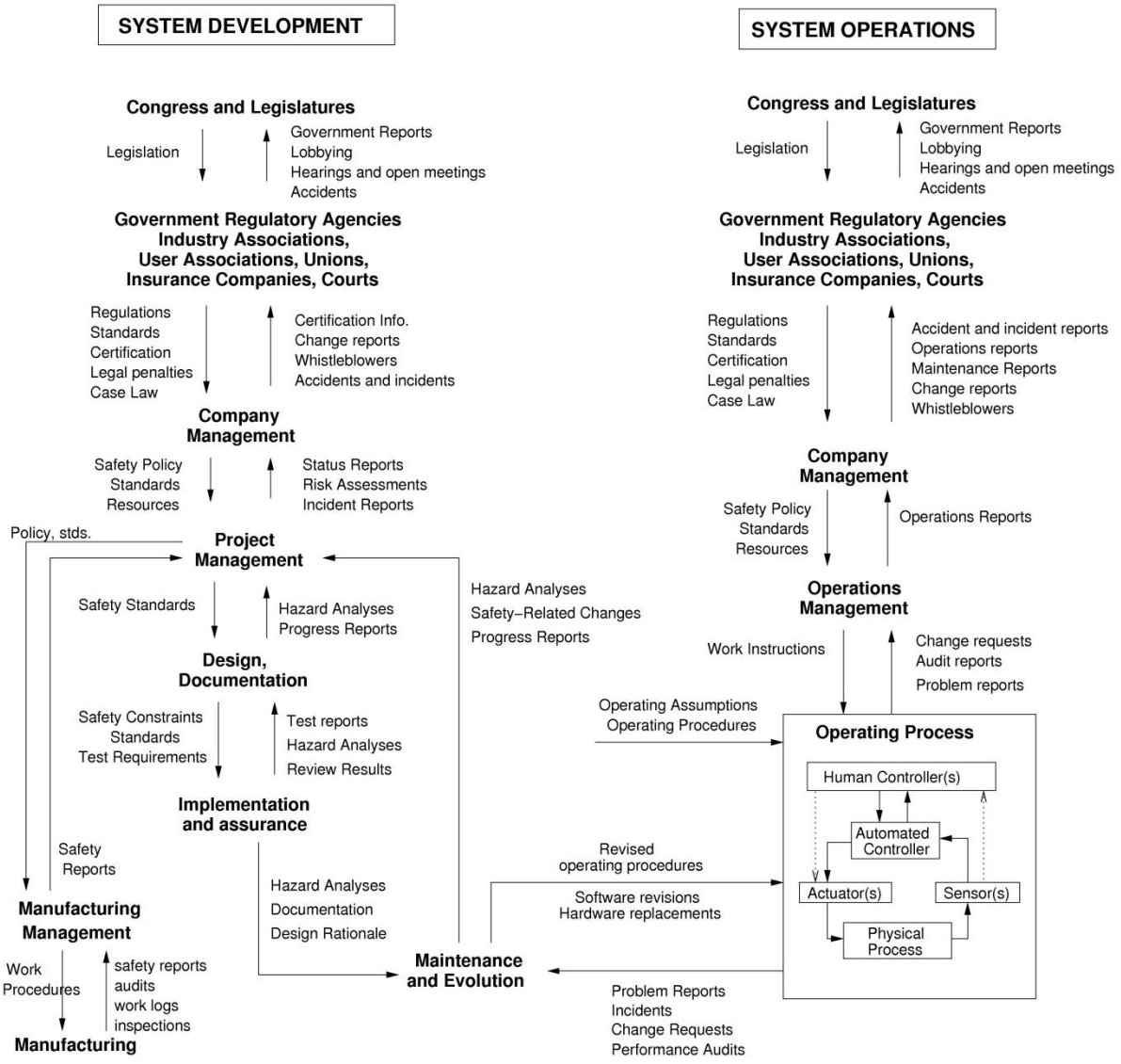


Figure 17: An example of a safety control structure

Notice that the operating process (the focus of most hazard analysis) in the lower right of the figure makes up only a small part of the safety control structure. There are two basic hierarchical control

structures shown in this particular example—one for system development (on the left) and one for system operation (on the right)—with interactions between them.[14]

Each level of the structure contains controllers with responsibility for control of the interactions between and behavior of the level below. Higher level controllers may provide overall safety policy, standards, and procedures (downward arrows), and get feedback (upward arrows) about their effect in various types of reports, including incident and accident reports. The feedback provides the ability to learn and to improve the effectiveness of the safety controls.

There are usually many interactions between the control structures, only a few are shown in the figure for readability purposes. Manufacturers must communicate to their customers the assumptions about the operational environment in which the original safety analysis was based, e.g., maintenance quality and procedures, as well as information about safe operating procedures. The operational environment, in turn, provides feedback to the manufacturer and potentially others, such as governmental authorities, about the performance of the product during operations. Each component in the hierarchical safety control structure has responsibilities for enforcing safety constraints appropriate for that component, and together these responsibilities should result in enforcement of the overall system safety constraints. Note that this is only one example of a safety control structure. Other examples may look very different and include a different number and type of hierarchical control structures.

An important difference between Systems Theory and the standard linear causality models is that more types of causality are included. Figure 18 shows three additional types of causality common in complex systems.

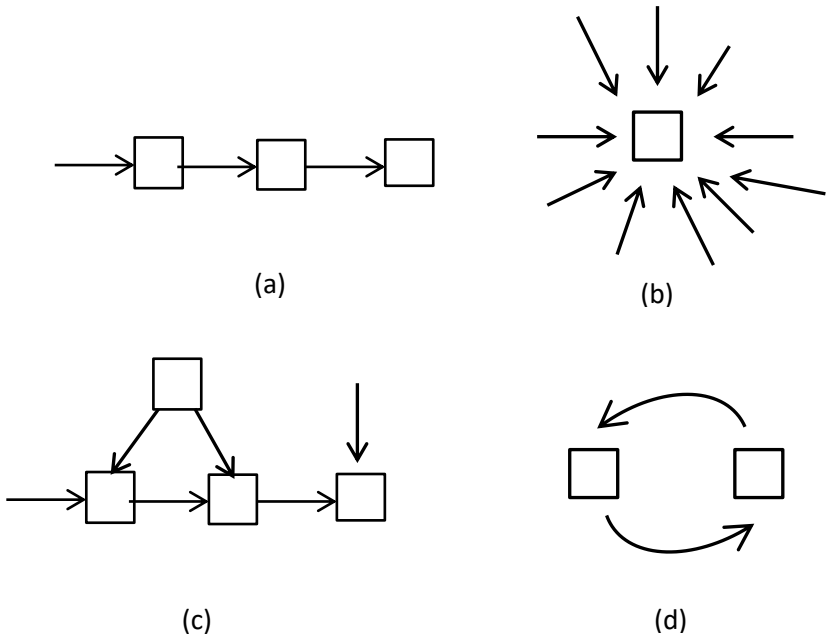


Figure 18: Four types of causality included in Systems Theory

[14] Prof. Hollnagel, on page 96, seems to assume (perhaps because he has seen this example) that there are always two, and only two, hierarchical control structures. That is not true. For example, Figure 17 might include a separate control structure for local government control of public health and emergency management in the case of an accident or multiple organizational or government structures.

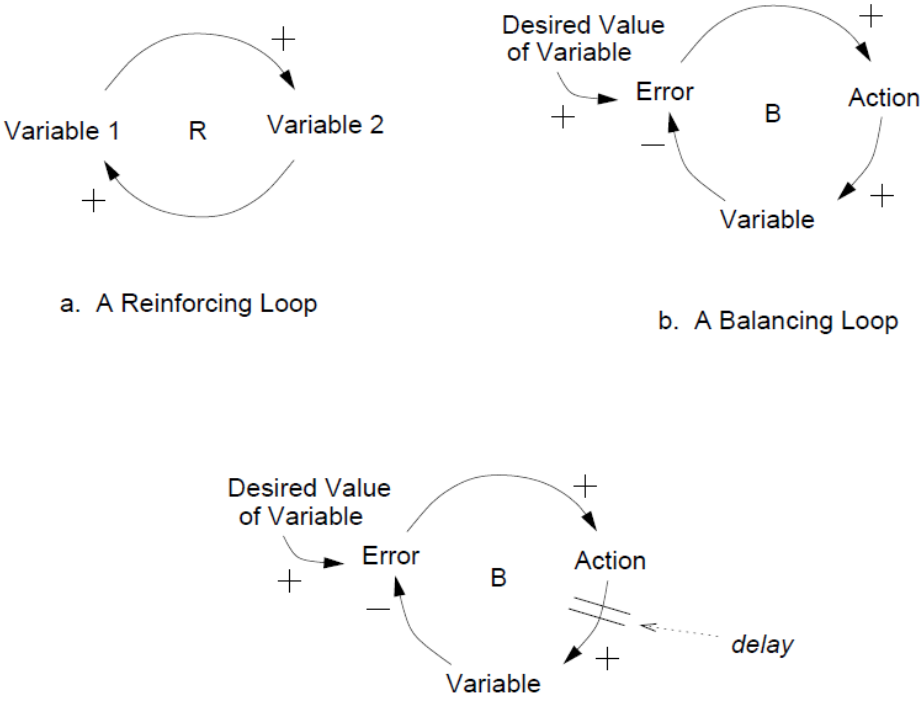
In addition to linear chains (shown in Figure 18.a), which is too simple a model to explain most real world behavior, Figure 18 shows how multiple independent "causes" can produce an effect or event (Figure 18.b), an event or condition can impact multiple events (Figure 1.c), and causality may be in the form of causal loops (Figure 18.d).

In Figure 18.c, there may be multiple common (systemic) factors that affect all the events. For example, some systemic factors (e.g., budget cuts or efforts to increase productivity) can defeat all the barriers at once.

A causal loop (Figure 18.d) incorporates feedback or feedforward relationships between events leading to an accident. The causal loop relationship between two events shown in Figure 18.d describes the possibility of multiple causes leading to a loss without a simple linear relationship between the events and causes. Causal loops are a way to model the dynamic changes in systems that lead to accidents. Event chains treat a system as a static, unchanging structure. But systems continually experience change and adapt to existing conditions. In contrast to the usually simple and direct causation represented in event-chain accident models, most accidents in complex, sociotechnical systems involve relationships between events that involve multiple feedback loops. Causal loops provide a framework for dealing with dynamic complexity, where cause and effect are not related in a simple way.

Figure 19 illustrates three basic causal loop structures: positive feedback or reinforcing loops, negative feedback or balancing loops (which tend to counteract change), and delays. Delays introduce potential instability into systems.

Figure 19.a shows a reinforcing loop, which is a structure that feeds on itself to produce growth or decline. Reinforcing loops correspond to positive feedback loops in control theory. An increase in variable 1 leads to an increase in variable 2 (as indicated by the "+" sign), which leads to an increase in variable 1, and so on. The "+" does not mean that the variables necessarily increase, only that variable 1 and variable 2 will change in the same direction. If variable 1 decreases, then variable 2 will decrease. A "–" means that they change in opposite directions.



a. A Reinforcing Loop

b. A Balancing Loop

Figure 19: Three types of causal loop structures

A balancing loop (Figure 19.b) changes the current value of a system variable through some action. It corresponds to a negative feedback loop in control theory. The difference between the current value and the desired value is perceived as an error. An action proportional to the error is taken to decrease the error so that, over time, the current value approaches the desired value.

The third basic element is a delay that is used to model the time that elapses between cause and effect. Delays make it difficult to link cause and effect (dynamic complexity) and may result in unstable system behavior. For example, in steering a ship, there is a delay between the change in the rudder position and a corresponding course change, often leading to overcorrection and instability. In aircraft, this phenomenon is called PIO or Pilot Induced Oscillation.

As an example of causality described using causal loops, consider a model of the cause of the Columbia Space Shuttle loss shown in Figure 20. A chain of events could be used to describe this accident (insulating foam flies off of external tank, hits and damages the thermal tiles, etc.). Indeed, any accident can be described using a chain of events, but it greatly oversimplifies the causal explanation.

The Space Shuttle was part of a larger Space Transportation System concept that arose in the 1960's when Apollo was in development. The concept originally included a manned Mars expedition, a space station in lunar orbit, and an Earth-orbiting station serviced by a reusable ferry, or Space Shuttle. The funding required for this large an effort, on the order of that provided for Apollo, never materialized, and the concept was scaled back until the reusable Space Shuttle, earlier only the transport element of a broad transportation system, became the focus of NASA's efforts. In addition, to maintain its funding, the Shuttle had to be sold as performing a large number of tasks, including launching and servicing satellites, which required compromises in the design. The compromises contributed to a design that was more inherently risky than was necessary. None of these factors are included in the model of causation shown in Figure 20 simply because of the need for an example in this paper that is limited in its complexity.

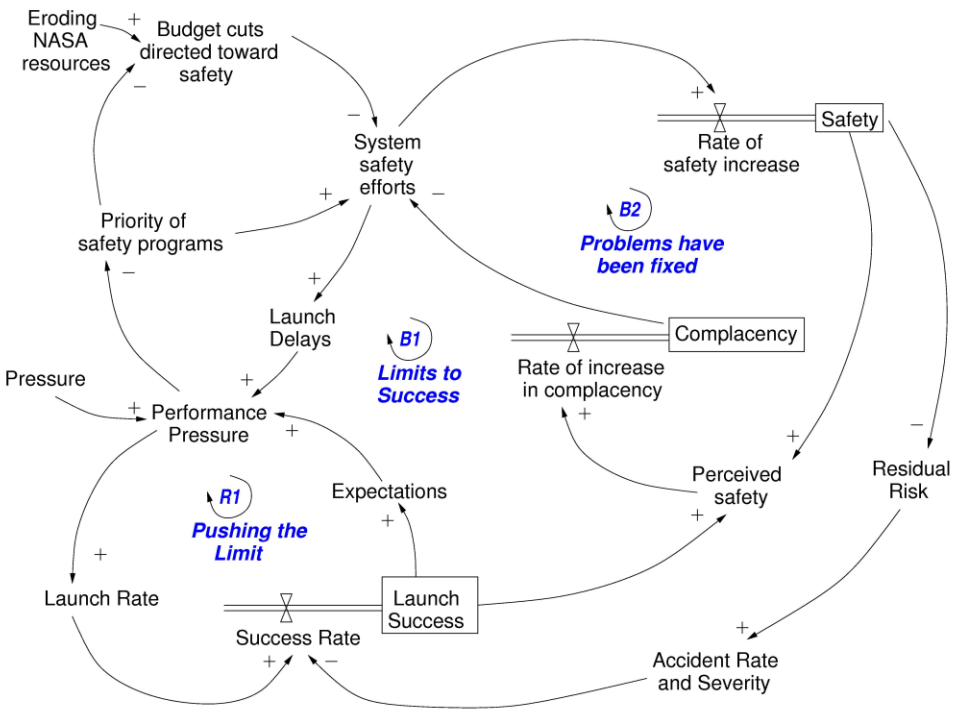


Figure 20. Some of the factors in the Space Shuttle Columbia accident:
The unrelenting pressure to meet the demands of an accelerating flight schedule

NASA also had to make promises about performance (number of launches per year) and cost per launch that were unrealistic. An important factor in both accidents was the pressures exerted on NASA by an unrealistic flight schedule with inadequate resources and by commitments to customers. The nation's reliance on the Shuttle as its principal space launch capability, which NASA sold in order to get the money to build the Shuttle, created a relentless pressure on NASA to increase the flight rate to the originally promised 24 missions a year.

Budget pressures added to the performance pressures. Budget cuts occurred during the life of the Shuttle, for example amounting to a 40% reduction in purchasing power over the decade before the Columbia loss. At the same time, the budget was occasionally raided by NASA itself to make up for overruns in the International Space Station program. The later budget cuts came at a time when the Shuttle was aging and costs were actually increasing. The infrastructure, much of which dated back to the Apollo era, was falling apart before the Columbia accident. In the past 15 years of the Shuttle program, uncertainty about how long the Shuttle would fly added to the pressures to delay safety upgrades and improvements to the Shuttle program infrastructure.

The budget cuts without concomitant cuts in goals led to trying to do too much with too little. NASA's response to its budget cuts was to defer upgrades and to attempt to increase productivity and efficiency rather than eliminate any major programs. By 2001, an experienced observer of the space program described the Shuttle workforce as "The Few, the Tired".[15]

NASA Shuttle management also had a belief that less safety, reliability, and quality assurance activity would be required during routine Shuttle operations. Therefore, after the successful completion of the orbital test phase and the declaration of the Shuttle as "operational," several safety, reliability, and quality assurance groups were reorganized and reduced in size. Some safety panels, which were providing safety review, went out of existence entirely or were merged.

The causal loop diagram shown in Figure 20 represents some of these factors.[16] The control loop in the lower left corner of the model, labeled R1 or *Pushing the Limit*, shows how as external pressures increased, performance pressure increased, which led to increased launch rates and success, which in turn led to increased expectations and increasing performance pressures. The larger loop B1 is labeled *Limits to Success* and explains how the performance pressures led to failure. The upper left loop represents part of the safety program factors in the accident. The external influences of budget cuts and increasing performance pressures reduced the priority of system safety practices and led to a decrease in system safety efforts.

The safety efforts also led to launch delays, which produced increased performance pressures and more incentive to reduce the safety efforts. At the same time, problems were being detected and fixed, which led to a belief that *all* the problems would be detected and fixed (and that the most important ones had been) as depicted in loop B2 labeled *Problems have been Fixed*. The combination of the decrease in system safety program priority leading to budget cuts in the safety activities along with the complacency denoted in loop B2, which also contributed to the reduction of system safety efforts, eventually led to a situation of (unrecognized) high risk where despite effort by the operations workforce, an accident became almost inevitable.

One thing not shown in this simplified model is that delays can occur along the arrows in the loops. While reduction in safety efforts and lower prioritization of safety concerns may eventually lead to accidents, accidents do not occur for a while so false confidence is created that the reductions are

[15] Harold Gehman (Chair), *Columbia Accident Investigation Report*, U.S. GAO, August 2003.

[16] Many other factors are contained in the complete model, including system safety status and prestige; shuttle aging and maintenance; system safety resource allocation; learning from incidents; system safety knowledge, skills, and staffing; and management perceptions [Leveson, 2006].

having no impact on safety. Pressures increase to reduce the safety program activities and priority even further as the external performance and budget pressures mount, leading almost inevitably to a major accident.

The STAMP Model of Accident Causality

To apply Systems Theory to safety, a new accident causality model is required that extends what is currently used. STAMP (System-Theoretic Accident Model and Processes) is based on Systems Theory and expands the traditional model of causality beyond a chain of directly-related failure events or component failures to include more complex processes and unsafe interactions among system components [Leveson, 2012]. In STAMP, safety and other emergent properties, such as security, are treated as a dynamic control problem rather than a failure prevention problem. Causal loops are modeled by standard engineering feedback control loops. No causes are omitted from the STAMP model that appear in the older causality models, but more are included and the emphasis changes from preventing failures to enforcing constraints on system behavior. All the factors shown in Figure 20 plus a lot more would be included in an analysis of this accident using CAST (Causal Analysis using Systems Theory), which is an accident analysis method based on the STAMP model of accident causality.

Some advantages of using STAMP are that:

- It can be applied to very complex systems because it models them top-down rather than bottom up. All the details of the system need not be considered, i.e., the system is treated as a whole and not as interacting components.
- It includes software, humans, organizations, safety culture, etc. as causal factors in accidents and other types of losses without having to treat them differently or separately. In other words, it handles true sociotechnical systems.
- It includes much more than simply failure events and the design techniques used to try to prevent them.

Using STAMP as the foundation, we have created tools for hazard analysis (STPA), accident analysis (CAST), identification and management of leading indicators of increasing risk, organizational risk analysis, etc. These tools have been shown to be much more powerful and effective in both analytical and empirical evaluations. The tools are starting to be widely used in some types of new automated and autonomous systems such as automobiles,[17] aircraft, and defense systems.

In STAMP, hazards are controlled by the hierarchical control structure such as the example shown in Figure 17. Note that the use of the term "control" does not imply a rigid command and control structure. Behavior is controlled not only by engineered systems and direct management intervention, but also indirectly by policies, procedures, shared value systems, and other aspects of the organizational culture. All behavior is influenced and at least partially "controlled" by the social and organizational context in which the behavior occurs. Engineering this social and organizational context can be an effective way to create and change a safety culture, i.e., the subset of organizational or social culture that reflects the general attitude about and approaches to safety by the participants in the organization or society [13].

Formal modeling and analysis of safety must include these social and organizational factors and cannot be effective if it focuses only on the technical aspects of the system. As we have learned from major accidents, managerial and organizational factors and often governmental controls (or lack of them) are as important as technical factors in accident causation and prevention. The technical, human,

[17] Cars today contain on the order of one hundred million lines of software. There are not enough examples yet to get the same estimates for autonomous autos, but the amount of software and complexity in these autonomous systems will surely not be less.

and social/organizational parts of a complex system are included in the STAMP control models and the analysis tools based on STAMP.

For space reasons, Figure 17 emphasizes the high-level components of the example safety control structure and not their detailed design, which can be quite complex. For example, the operating process (lower-right-hand box) might include all the physical and control parts of an aircraft or a nuclear power plant.

Figure 21 shows the basic form of the interactions between the levels of the control structure, where the controller imposes control actions on the controlled process. The standard requirements for effective management—assignment of responsibility, authority, and accountability—are part of the control structure design and specification, so both hardware, software, and human controllers are included, as well as important aspects of management control.

The importance of feedback becomes apparent here. In engineering, every controller must contain a model of the controlled process in order to provide effective control. For human controllers, this model is usually called a *mental model.* This process model or mental model includes assumptions about how the controlled process operates and the current state of the controlled process. It is used to determine what control actions are necessary to keep the system operating effectively.

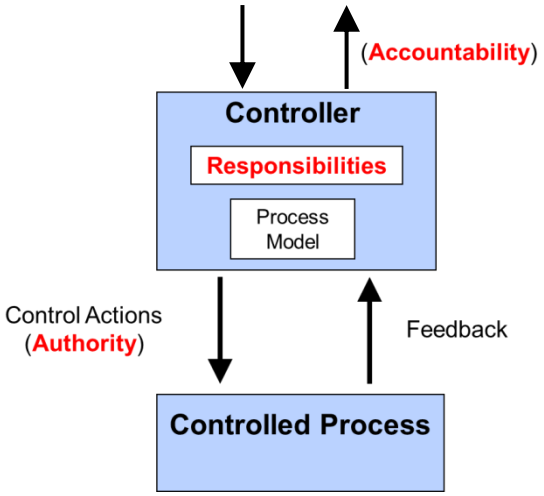


Figure 21. The basic building block for a safety control structure

Accidents in complex systems often result from inconsistencies between the model of the process used by the controller and the actual controlled process state, resulting in the controller providing unsafe control actions. The controllers may be automated or human. The controllers, either human or automated, are assumed to continually vary their behavior and adapt to the current circumstances. Note that this variance allows much more complex behavior than the simple performance variability due to approximate adjustments described by Hollnagel in his Resonance model although there is clearly a relationship between the two.

As examples of unsafe behavior, the autopilot software may think the aircraft is stalled (an incorrect process model) and issue an order to descend when in reality the aircraft is not stalled. The pilots, surprised by this aircraft movement, may respond and try to ascend. This interaction may end up moving the aircraft into an unsafe state. A military pilot may think a friendly aircraft is hostile and shoot a missile at it. The software controlling a spacecraft may think the spacecraft has landed and turn off the descent engines prematurely. As a final example, the early warning system may think the country has

been targeted by an enemy missile and launch a counter attack. All of these examples have already occurred or, in the case of the missile attack, came close.

Note that it does not matter whether the incorrect process model was a result of an unintentional or intentional cause so that security can be handled in the same way. The Stuxnet worm in the Iranian reactor program is an example. The worm changed the controller's process model to think that the centrifuges were spinning slower than they were and the controller reacted by sending "increase speed" commands to the centrifuges, wearing them out prematurely.

Part of the challenge in designing an effective safety control structure is providing the feedback and inputs necessary to keep the controller's model of the controlled process consistent with the actual state of the controlled process so that only safe control actions are issued.

There may be a large number of controllers and control loops. The two structures in Figure 17 are an artifact only of the example shown. Note again that this is not some draconian, authoritarian, dystopian view of safety. It is, in fact, exactly the way the world and safety management works today. It is simply a more complete and realistic model of accident causality than other common ones based on linear chains of events. Unfortunately, it does not lend itself to simple diagrams showing causality using dominoes or Swiss cheese slices. The best I have been able to come up with is shown in Figure 22.

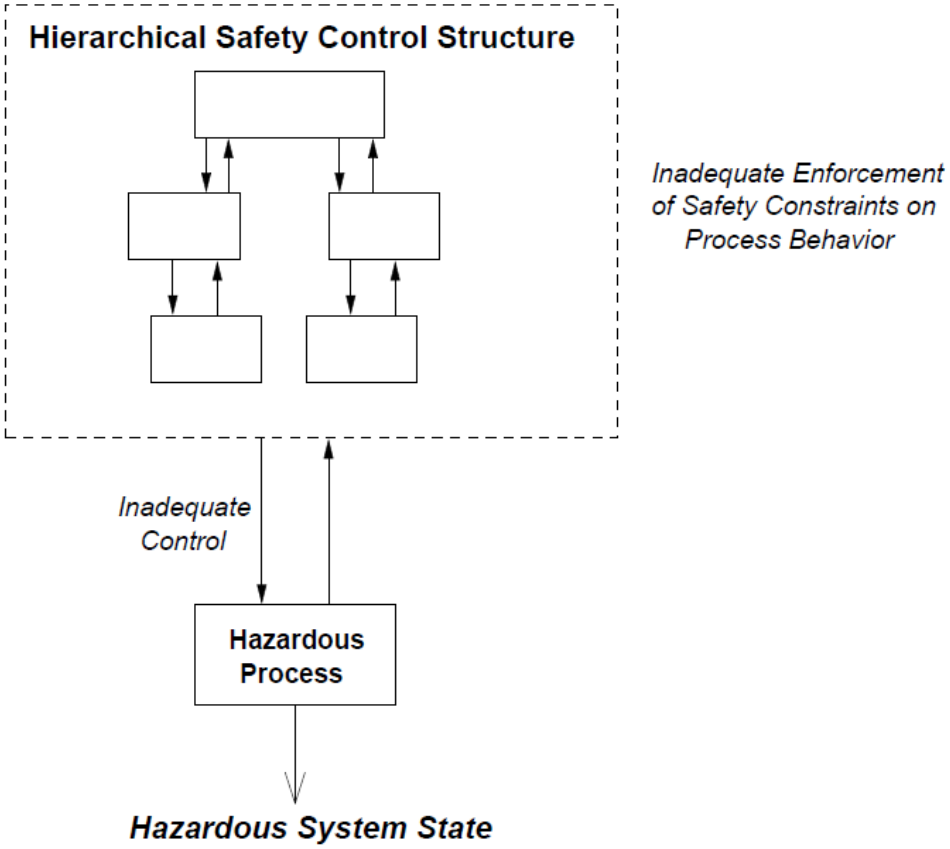


Figure 22: A representation of the STAMP model of accident causality

To summarize, Safety-I, Safety-II, and most traditional safety engineering use a linear causality model. While an epidemiological model was proposed in 1948, it has had little impact and is likely not practical. STAMP is a new, more powerful and inclusive model of causality that underlies the approach labeled here as Safety-III.

Attitude Toward Human Factors

	Safety-I	Safety-II	System Safety Engineering	Safety-III
Attitude to the human factor	Humans are predominantly seen as a liability or a hazard.	Humans are seen as a resource necessary for system flexibility and resilience.	Humans are expected to prevent or respond to hazards and to be flexible and resourceful when they occur.	The system must be designed to allow humans to be flexible and resilient and to handle unexpected events.

Prof. Hollnagel uses the term "humans" a lot, but put into context, it appears that he is talking almost solely about low-level workers and operators in simple workplaces. There are lots of humans in our systems; humans design systems, manage them, certify them, maintain them, as well as operate them. But most of his assertions seem only to apply to operators and seem to omit the sophisticated human factors research and human-centered design in engineered, complex systems today.

> Page 120: *During the history of system safety, the human factor has been considered first as a bottleneck that hindered the full use of the technology potential.*

Prof. Hollnagel must be referring to workplace safety here as this statement has never been true of product/system safety. Human operators were used in system safety as a way to *achieve* full "technology potential" when all of the functions could not be automated. We have not yet tried to automate managers, certifiers, government oversight agency personnel, or designers (although designers are using more automated support tools).

> *and then as a liability or threat that not only limited performance but was also a source of risk and failure ('human error')*.

In product/system engineering, much more emphasis has always been placed on the hardware failures. It is true, however, that users and operators have been a convenient scapegoat after accidents for mistakes made in engineering and management. Given that there are no citations or examples, his statement above appears to be another strawman argument.

Elsewhere, Prof. Hollnagel argues that

> *And since a human is usually in a system instead of a piece of equipment because it is cheaper or because we might not know how to construct a mechanical component with the same performance, all we need to do is to make sure that the human functions like this component."*

Human operators are not cheaper than most equipment. For example, airline executives would love to replace pilots with mechanical equipment because pilots are in fact very expensive. The same is true for human operators in most industries. Engineers don't keep humans in systems because they are cheaper but because they are able to vary their performance exactly the way that Prof. Hollnagel is suggesting they do. So it would be very strange for engineers to expect that humans operate like physical devices; this would be irrational. And, of course, Prof. Hollnagel only considers human operators and not even other humans in the system. Nobody is considering replacing CEOs or executive management with automation because the automation is cheaper (although, in the case of executive management remuneration today, it sounds like a good way to greatly decrease expenses if it were possible).

> *But by recognizing the necessity of and value of performance variability, the human factor becomes an asset and a sine qua non for system safety.*

In fact, this is exactly what is done today. Prof. Hollnagel is again arguing against a strawman.

Prof. Hollnagel seems to believe that Taylorism, which he described in detail in his book and unfortunately too often sometimes still has influence in assembly line manufacturing, exists or ever existed in product/system safety. Humans have always been considered an asset in product/system safety.[18] Again, that's why they are left in systems today. In fact, if the operator role could be described using procedures that we wanted operators to comply strictly with, we would simply automate the functions and dispense with human operators completely. Where we need repeatability (the catchword in Taylorism), functions are automated. We leave humans in systems precisely because we *need* them to provide variability and to deal with unexpected events. We depend on operators to do what cannot be definitively defined as always the safe thing to do in all situations. This is not a new idea that Prof. Hollnagel has invented. His characterization of Safety-I seems to apply only in workplace safety, if anywhere.

> *This change in perception has not come about abruptly, however, but has been slowly growing for the last 15 years.*

This perception has always been true in engineering so I am not sure in what field he thinks the change has come about. Workplace safety? Certainly not product/system safety. And healthcare has always depended on doctors and nurses; extensive automation is fairly recent.

Computers and humans both provide control in advanced automation today. One of the difficult questions engineers must face is who should be given the final authority−human operators or automation. Some accidents have occurred because the human operators were overruled and unable to do what was necessary to prevent an accident because the computers were confused about the actual state of the aircraft. For example, the computers think the aircraft is in the air and do not allow the pilots to activate the reverse thrusters, as described earlier in the Warsaw A320 accident. This is the very difficult question that engineers in systems with shared human and automation control face today: who should be given the final authority for action? The problems we are struggling with are not the simple questions that Prof. Hollnagel raises—we've already tackled those.

If we want resilience in systems, then it needs to be integrated into the design of the entire system. Human operators rarely have the ability to provide resilience in highly automated systems unless the engineers have designed in the ability for them to provide it. What is needed, which I call Safety-III, is described in more detail later in this paper.

What is most surprising to me about these two books is that Prof. Hollnagel seems to think he is taking a "systems approach" when in fact he is doing the exact opposite. The books are laser focused on only one small part of the sociotechnical system, the small box in the example safety management system shown in Figure 17 in the lower right hand corner labeled "human controller" in the box labeled "Operating Process." Direct human operation (control), as stated above, is not the only role that humans play in systems. For the most part today, humans are not directly controlling processes but controlling computers that actually control the physical process.

Prof. Hollnagel continues and uses another strawman argument:

> *To be consistent with this role, accident analysis and risk assessment must acknowledge the following:*
>
> - *Systems are not flawless and people must learn to identify design flaws and functional glitches.*

Having done hazard analysis and risk assessment for 40 years, I cannot imagine anyone who would *not* acknowledge this. In fact, that is exactly the role of hazard analysis, accident analysis, and risk assessment, i.e., to identify design flaws and functional glitches, that is, how flawed system behavior can

[18] One exception perhaps is recently in totally automated cars, but it is unclear how successful this effort will be.

lead to accidents. If those doing accident analysis truly believed that systems were flawless, then why would they bother doing it? The *goal* of accident analysis is to look for flaws in systems.

What "people" could Prof. Hollnagel be referring to here? It certainly could not be engineers in general or system safety engineers in particular. Engineers who assume that their physical systems are going to work ideally will soon be out of a job. The important fields of reliability engineering and system safety are based on the assumption that systems are not flawless. If the "people" he is referring to are operators, remember that operators are part of the system, and they too are not flawless.

Engineers always expect that systems might not work perfectly. They do their best to make them work correctly and then try to reduce erroneous or unsafe behavior through design. They design systems with human operators to deal with the events that result from the system not working flawlessly. The fact that people can adjust work to the actual conditions is exactly why we don't automate them out of system designs.

- *People are able to recognize the actual demands and can adjust their performance accordingly.*

Again, it is not clear who "people" are. As explained above, humans are left in systems today to do exactly this. If by "people" Prof. Hollnagel means human operators, then, unfortunately, the statement that human operators can recognize actual demands and adjust performance accordingly is not always true. Their ability to do this depends on the design of the larger system, not only on their desire to do so. As an example, healthcare workers may want to be flexible but are often hindered in this desire of the design of electronic health systems.

In Safety-III, systems are designed so that operators can be resilient and do the things that Prof. Hollnagel and everyone else wants them to do. An earlier example cited the accident report for the American Airlines B757 crash while landing at Cali, Columbia that concluded one of the causes was the flight crew being confused by automation that was confusing and demanded an excessive workload rather than a conclusion that the cause of the accident was the flawed automation design. This type of shortsightedness is what happens when we rely too much on the human operator to be able to adapt to any situation. If engineers provide poor tools for them to use, then the operators (pilots in this case) may not be able to avoid accidents.

- *When procedures must be applied, people can interpret and apply them to match the conditions.*

In fact, again this is exactly what we expect from them. I realize that in many poor instances of workplace safety that procedure compliance is expected. But product/system safety is not workplace safety; Taylorism never played any role in engineering.

There are many instances, however, when procedures are useful and compliance important. Pilots, for example, cannot be expected to be experts on and understand everything about the design of the aircraft. Even many engineers are only experts on parts of the design of very complex systems. In addition, pilots usually do not have the time to start reasoning from first principles. So they are given procedures to apply based on the current conditions and when they do not have time to read through a manual to find a solution to a problem. If the procedures do not work, then something else has to be done, and we try to train pilots and depend on them to make good decisions in these cases. But that does not mean that we should not train the pilots on procedures and when it is appropriate to apply those procedures. Flying would be a lot more dangerous if we did not. We don't want all human operators to have to learn from first principles or experience each time they have to make decisions. In more everyday examples, we also depend on skills or learned procedures when time is of the essence. If I had to figure out what to do every time I have to stop my car to avoid an accident and not just automatically step on the brakes (a learned procedure), I would be involved in a lot of accidents.

Page 45: *According to this way of looking at the world, the logical consequence is to reduce or eliminate performance variability either by standardizing work, in the spirit of Scientific Management* [Taylorism] *theory, or by constraining all kinds of performance variability so that efficiency can be maintained and malfunctions and failures avoided.*

Taylorism never had any influence on engineering. I cannot understand where Prof. Hollnagel got this impression. His books contain no references or citations that prove that product/system engineering has "this way of looking at the world." Do humans "fail" or "malfunction"? Or is he suggesting that performance variability by human operators prevents hardware and software from failing and malfunctioning?

Page 45: *Performance can be constrained in many different ways, some more obvious than other. Examples are rigorous training (drills), barriers and interlocks of various kinds, guidelines and procedures, standardization of data and interfaces, supervision, rules and regulations.*

I want my pilots, physicians, etc. to undergo training. Barriers and interlocks are used to prevent the system from getting into a hazardous state, i.e., a barrier across railroad tracks when a train is approaching and the automobile driver cannot or is likely not to detect it in time. Standardization of data and interfaces is used to reduce errors and make systems easier to operate and to learn how to operate. Rules and regulations are a necessity in any complex system or society in order to avoid chaos. Is Prof. Hollnagel really suggesting that we get rid of all these things? Of course, not all rules and regulations or barriers are good. But it is not a black and white issue, all bad or all good. They are often important in preventing specific types of dangerous performance variation.

To take a healthcare example, electronic prescription systems will usually detect when a physician appears to be ordering a prescription for a lethal dosage or in the wrong units—perhaps because of a typo—or is ordering a prescription that has a bad reaction with another prescription that patient is taking. Should we really get rid of this type of protection? The response may simply be to ask the physician if that is what they really want rather than disallowing it. Prof. Hollnagel seems to have a very narrow view of what types of protection can be used. The world would be a much more dangerous place if all the protections that have been designed are removed in order to "not constrain human performance."

- *People can detect and correct when something goes wrong or when it is about to go wrong, and hence intervene before the situation seriously weakens.*

This statement is only true under two conditions. The first is that the system has been designed to give them the information they need to detect when something has gone wrong and to intervene successfully and in a timely manner. For example, there have been accidents and incidents where an aircraft engine was failing but the autopilot compensated for it. The pilots only learned about the failing engine when the autopilot finally gave up, leaving the pilots with inadequate time to diagnose the problem and react.

The second necessary condition is that the system has been designed to provide the operator with the ability to intervene. If the operators are not provided with direct controls in an automation-intensive system, the humans may be adaptable and resilient but will not have the tools to prevent losses. If a doctor knows that the standard medication for treating a patient is worse than an alternative in a particular case, the physician must have the ability to obtain the alternative medication. That is a system design problem, not just a function of the adaptability of the physician.

Page 138*: Unacceptable outcomes or failures cannot be prevented by eliminating or constraining performance variability since that would also affect the desired acceptable outcomes.*

This is another black and white statement and overgeneralization. There are times when constraining performance variability is exactly the right thing to do even though it could affect desired acceptable

outcomes. We require that pilots lock the cockpit door in the aircraft in order to keep out those wanting to do harm to the aircraft and passengers. There are probably some odd cases where it might be better not to have cockpit doors locked (as in the case of the pilot suicide in the Germanwings aircraft). But those cases are much less likely than the need to prevent entry to bad guys. These are the types of difficult decisions that must be made in engineering. How much and what kind of protection do we provide to prevent unintentional or intentional behavior of operators? This is one of the most difficult decisions made in safety engineering. There are *always* tradeoffs.

> *Instead, efforts are needed to support the necessary improvisations and performance adjustments by understanding why and how they are made, by clearly representing the resources and constraints of a situation, and by making it easier to anticipate the consequences of actions.*

It would be hard to find anyone who would disagree with this statement. But Prof. Hollnagel is simply describing some basic principles in the field of human factors engineering. They have existed for a very long time. This is not a novel approach. And it is pretty much standard practice in engineering.

At the same time, it is a great deal more difficult to do than Prof. Hollnagel implies. For example, making it easier for operators to anticipate the consequences of their actions has been suggested and implemented in *predictor displays*. Back in 1981, for example, Moray argued that operators should be given an estimate of the future states of the system. Determining the consequences of actions in real-time, however, can be quite difficult in a complex system. If we could always do that, we could probably design the system to always avoid hazard states using automation. There are also many drawbacks, such as operators reducing vigilance and putting too much confidence in automated predictions or automated assistance. This subject is too complicated to go into here, but Prof. Hollnagel is ignoring that such efforts exist and have existed for a long time and also underestimating the difficulty in accomplishing these goals. If they have not been totally achieved, it is not because people have not tried.

And, to be knowingly repetitive, the system must be designed to prevent the improvisations and adjustments that will lead to hazards, such as by adding positive train control. There are always advantages and disadvantages of either adding prevention measures in the physical design or assigning all responsibility to the human operators.

> *Performance variability should be managed and dampened if it looks like it is going in the wrong direction or amplified if it looks like it is going in the right direction. In order to do this, it is necessary first to recognize performance variability, second to monitor it, and third to control it.*

This goal is easily stated, and again something has been tried for decades, but very hard to achieve successfully. One can put in physical barriers to prevent some unsafe variability. An example is not allowing the system to be restarted if human maintenance actions have not been completed or if automated protection systems have not been restarted. In engineering, these are called *interlocks* and have been used for at least a hundred years. Physical interlocks are now often replaced by software interlocks that try to prevent unsafe operator behavior, exactly as Prof. Hollnagel suggested above. But whenever we design systems in this way, we may also prevent situations that we want to happen. For example, saving an aircraft from crashing may require taking actions that could damage the aircraft.

Prof. Hollnagel's quote above implies that it is possible to determine what humans will do in the future and determine whether their actions are going in the "wrong direction" or in the "right direction." The problem again goes back to Prof. Hollnagel's use of undefined and vague terms like "wrong direction" and "right direction." Not only are they undefined, but they assume there is always one "right" direction and one "wrong" direction and no difficult tradeoffs are involved. Once again, it is very easy to suggest and very difficult to do.

In addition, we do have monitors, but how does one know the monitor is correct? In the B737 MAX accidents, the software thought it was doing the right thing. Should it have prevented the pilots from

taking a counter action? The Warsaw A320 accident described previously is another example, where the software did not allow the operators to operate the reverse thrusters to slow the aircraft after landing because it thought the pilots were doing the wrong thing. Who has final authority? If we could always determine what was the "right" or "wrong" direction for behavior, then we would, again, be able to automate the actions and take the human operators out of the system. Who guards the guards—*quis custodiet ipsos custodes*?

In fact, if we could always determine if what the operator was doing was right or wrong, we would not need human operators at all. Once again, Prof. Hollnagel oversimplifies the problems and ignores decades of research and progress on these topics.

In engineering, humans have been expected to prevent or respond to hazards and to be flexible and resourceful when they occur. Note the use of the word "hazard" in the previous sentence, which is defined, as opposed to undefined terminology such as "right" and "wrong."

In Safety-III, it is assumed that human ability to be resilient and to respond may be limited by the design of the system in which they are working. Therefore, resilient *system design* needs to be the goal, not just resilient humans. The most critical need is to learn how to design systems in which humans are able to be resilient. And, of course, the entire system must be resilient, not just the human operators.

Role of Performance Variability

	Safety-I	Safety-II	System Safety Engineering	Safety-III
Role of performance variability	Harmful, should be prevented as far as possible.	Inevitable but also useful. Should be monitored and managed.	The primary reason for performance variability is to enhance productivity and system requirements. Procedures are provided when response time and information is limited. Effort is put into providing appropriate controls and interfaces to allow operators to prevent or respond to hazards. Design so that when performance (of operators, hardware, software, managers, etc.) varies outside safe boundaries, safety is still maintained.	Design the system so that performance variability is safe and conflicts between productivity, achieving system goals, and safety are eliminated or minimized. Design so that when performance (of operators, hardware, software, managers, etc.) varies outside safe boundaries, safety is still maintained.

As noted above, the strawman Safety-I belief that performance variability is harmful and should be prevented as far as possible does not exist outside Taylorism, which was applied to simple workplaces and assembly line processing over a hundred years ago to increase productivity, *not* safety. Once again, his characterization of Safety-I does not exist and never did, except perhaps in workplace safety. This is probably why so much emphasis is placed in his books on the history and approaches of workplace safety, such as Heinrich, his model and his triangle, procedure compliance, and Taylorism. None of these was influential in product/system safety.

Prof. Hollnagel claims that in Safety-II, performance variability is inevitable but also useful. It should be monitored and managed. This is quite easy to say and quite difficult to actually do in system engineering design. Some of the limitations of our ability to do this were described above.

What about real product/system safety instead of the strawman Prof. Hollnagel has created? In general, the belief in engineering is that the primary reason for performance variability is to enhance productivity and achievement of system requirements. We count on the operators not to blindly follow procedures. Procedures are provided when response time and information is limited. Effort is made to provide appropriate controls and interfaces that allow operators to prevent or respond to hazards when procedures are not appropriate to use. Finally, a goal is to design the system so that when performance (of operators, hardware, software, managers, etc.) varies outside safe boundaries, safety is still maintained. That property, in engineering is called fault tolerance and fail-safe design.

In Safety-III, the goal is to design the system so that performance variability is safe and conflicts between productivity, achieving system goals, and safety are eliminated or minimized. In addition, the design should ensure that when the performance (of operators, hardware, software, managers, etc.) varies outside safe boundaries, safety is still maintained. More generally, resilience in Safety-III is defined as the ability of the system to maintain the safety constraints in the face of unplanned or inadequately controlled behavior or hazards. In a resilient system, losses are prevented or minimized in the face of unexpected events. To accomplish this goal, we need to design the entire sociotechnical system to be resilient, not just part of it and, especially, not just assume that the human operators will be able to provide it in any system design.

The important principle when taking a systems view of safety is that the system must be *designed to allow successful resilience by human operat*ors. There are three requirements to accomplish this goal:

1. The humans in the system, including operators, managers, government overseers, etc. must be aware that a hazardous state has occurred. The hazardous state must be observable within the time period necessary to prevent a loss. The same is true for software or any type of automated system controller that we expect to respond to hazards. Without knowing that a hazard exists, then it is not possible to control it or to respond to minimize damage.
2. Accurate information about the current state of the system must be available in a timely manner. The operator must have the information necessary to solve problems.
3. The system design must allow the flexibility required to be resilient. If there are no actions that the controller (human operator or manager, software, social structure) has available to recover from a hazard, then the human can be resilient—in terms of knowing what to do—but not be able to respond in any effective way. A simple example is the “undo” button in many software applications. More specifically, if a hazard does occur, perhaps because of errors on the part of the operators themselves, either they must have a means to reverse the errors or move to a non-hazardous state. Or other parts of the system must have this capability.

By using STAMP or some type of system theoretic modeling language and analysis tools built on it, a system can be analyzed to determine whether it truly incorporates the potential for resilience. And, even better, it can be designed from the beginning to be resilient. We cannot just assume that operators will be able to be resilient and prevent potentially enormous losses. At Fukushima, the ability of the

operators to shut down the nuclear reaction was thwarted by the fact that the electrical supply needed to do so was located in the basement, which was immediately flooded by the tsunami and became inoperable. The operators were potentially resilient, but the design prohibited them from taking effective action. This design flaw was potentially identifiable using appropriate hazard analysis techniques such as STPA.

Actually, none of this is new. Jens Rasmussen (who Prof. Hollnagel worked with at Riso Labs) in 1987 wrote about a concept he called design for error tolerance [Rasmussen, 1987a, 1987b, 1990]. He suggested that the ability of the operator to explore should be supported by the system design and that a means for recovery from the effects of errors should be provided. This goal can be achieved by the design of work conditions (I would generalize that to "the design of the system") in which errors are immediately observable and reversible.

Rasmussen described two system design requirements for error tolerance:

1. Providing Feedback: To make errors or hazardous states observable requires feedback or some way for the controllers to obtain accurate information about the current state of the system.
2. Allowing for Recovery: Controllers must have adequate flexibility to cope with undesired system behavior and not be constrained by inadequate control options. There must also be enough time for the recovery actions to be taken. For recovery to be possible, controllers (both human and automated) must be provided with compensating actions for recovery actions that are not successful and may themselves have undesired effects.

Note that human resilience requires that the system be designed to allow the human to be effective in handling unforeseen events. That sounds impossible, i.e., how does one design for something that is not foreseen? In fact, it may not be possible in all cases. But we can in many cases provide the human operator with the information about the current state and with ways to modify the state. Clearly, however, we cannot provide controllers (whether automated or human) with unlimited information or the means to control everything. What that means is that resilience will always be limited but better hazard analysis and design techniques may help us to stretch the limits.

Sometimes designers are so sure that their systems can correct themselves that they do not include appropriate means for human intervention. Clearly, systems can (and do) fail or, more generally, behave in ways that the designers did not anticipate. But if the designers do not provide appropriate means for the operator to intervene, losses can occur.

In the end, however, there are always limits to how much resilience can be designed into systems: human controllers, totally automated controllers, or even humans and automation working together will not be able to prevent or minimize losses in all cases. We need to keep that in mind when decisions are made about whether to build and operate extremely dangerous systems and how much autonomy to allow.

How do we create resilient systems? A (sociotechnical) system engineering process that integrates STAMP (system thinking) tools into the development and operations process has the potential for creating resilient systems. Simply focusing on what human operators do when there is no accident (i.e., when "things go right") is not adequate. We need to take a larger, systems perspective on the problem. My book, *Engineering a Safer World*, provides advice on how a systems approach to safety can allow us to achieve these goals.

In summary, nobody would argue against allowing performance variability in operators. The problem is that doing so is not as simple as monitoring and managing performance variation. We cannot assume that such variability will always be positive and that what is needed for it to be successful in a certain case will always be available without preplanning and careful system design.

In addition, performance variability is only a small part of what is needed to engineer/design sociotechnical systems to be safe. We cannot just focus on human operators or even on the "socio"

parts of systems. There must be an integrated design of the entire system for safety and other desirable properties. Performance variability is itself just a small part of even human factors engineering.

Finally, performance variability in all parts of the system (not just the humans) is much more difficult to tackle than implied in Prof. Hollnagel's books and certainly requires more than simply focusing on what humans do "right." The problem is that human operators cannot know whether their actions are safe or not until an accident or near miss occurs. An operator can consistently do something without an accident occurring simply because not all the conditions necessary for an accident have occurred yet. But the next day, the operator doing the same thing could lead to a catastrophe. Concentrating on when "things go right" is not going to provide the information necessary to engineer resilient and safe sociotechnical systems.

Summary

This paper is much longer than I hoped or expected. A summary is probably in order as a lot has been covered. One of the biggest problems is the faulty logic involved in Prof. Hollnagel's arguments. They are primarily of the form: Everyone does A (Safety-I) but doing B (Safety-II) would be better. The problem is that in most cases nobody is doing A. In some cases, they are already doing B. Most important, what is omitted from the argument is that A and B are not the only choices. There is a third choice C, which describes what is done almost universally in safety engineering today and has been done for at least 70 years.

Finally, there is a fourth choice D, here called Safety-III, which is different than A, B, or C, but its roots also go back 70 years in practice, primarily in defense and space systems. Recent advances to this traditional System Safety approach include a new model of accident causality based on Systems Theory called STAMP and the new analysis and design tools based on STAMP [Leveson, 2012].

To summarize what has been argued in this paper:

Safety-I: Doesn't exist as described.

Either Prof. Hollnagel is uninformed about safety engineering or he has created a strawman to make Safety-II look better. I know of no industry or safety application that is primarily reactive or relies on accident investigation as their primary way of improving safety, even healthcare and workplace safety. Everyone does, of course, investigate accidents to learn from them—they would be foolish not to do so. The other statements about Safety-I are equally uninformed, except for the part about linear causality models, which *are* prevalent today. Unfortunately, as best I can understand by the vague definitions and lack of examples I have found about "functional resonance" and FRAM, it appears that Safety-II also has an underlying linear causality model.

Safety-II: Concentrates almost entirely on the human operator.

The design of the system in which the human is operating seems to be ignored. Safety-II is the opposite of a systems approach (or a sociotechnical approach) as the technical seems to be excluded from playing any important role in safety. In reality, human flexibility and resilience is always limited by the overall system design system in which humans exist and work. Without considering that system but expecting the human to be resilient within it is unrealistic. Safety-II appears to be based on a linear chain-of-events model, pretty much identical to the Swiss cheese model except that other ways for the cheese holes (failure events) to be reached ("approximate adjustments") are hypothesized beyond simple failures. FRAM uses traditional functional decomposition.

Engineering for safety as actually practiced for the past century: Based on preventing hazards using a wide variety of methods.

In those industries that have been most successful in preventing accidents and where the most effort has been devoted to it, such as commercial aviation, a comprehensive safety engineering approach has been applied. Such a comprehensive approach includes modeling and hazard analysis, design for safety and fault-tolerance, safety management systems, human-factors engineering and human-centered design, safety programs during operations, regulation and licensing, event reporting systems, and, yes, accident investigation and analysis.

The least successful fields in reducing accidents, such as workplace safety and healthcare safety, have focused on the behavior of the workers and not on designing (engineering) the entire system for safety. But even these industries do try to take a proactive approach and are not just reactive or simply focus on after-the-fact accident investigation.

Safety-III: Based on the assumption that losses result from inadequate control of hazards.

Safety-III is based on Systems Theory. It spans the entire lifecycle but puts particular focus on designing safety in from the very beginning of system concept definition. Resilient systems are created not by simply focusing on human operators but instead by carefully designing the system to prevent and control hazards as much as possible, including the operator as a critical component of that design. In addition, if emergencies do arise that the system is not designed to handle, the system is designed so that human operators can be successful in handling it and that the tools exist for humans to be resilient in the case of emergencies.

Safety-III recognizes that change and adaptation to change are both inevitable and healthy in the system's lifetime. Planned changes must be carefully analyzed to ensure that they do not increase risk nor introduce new hazards. Appropriate procedures must be used to identify unplanned and hazardous changes, including changes in the external environment or in the assumptions upon which safety during design was based, and to provide an appropriate response.

Safety-III includes the design of sophisticated safety management systems where the desired safety culture is defined and nurtured, and there is a comprehensive and carefully designed safety management structure as well as a comprehensive and usable safety information system.

Safety-III is described in detail in my book *Engineering a Safer World* and the hundreds of papers that have been written on the approach suggested in the book and the analysis tools (including STPA and CAST) that allow implementing the approach. It is being used, to a greater or lesser degree, in every industry today.

Is Safety-III too expensive? As has been repeatedly shown through the past 100 years, productivity and safety go hand-in-hand for hazardous industries and products. They are not conflicting except in the short term. In the long term, engineering and designing for safety is a key to success and profits.

The Future

I promised at the beginning of this long paper that I would include where I think we need to go in the future. Here are some of the directions, in my experience, that would be most helpful:

- Implement Safety-III, i.e., a systems approach to safety, in all fields, including healthcare and workplace safety.
- Continue to improve and extend Safety-III. Create new holistic hazard/risk modeling and analysis techniques that include all facets of the sociotechnical system and how they can operate together to prevent losses. These tools should assist in making difficult conflicts and tradeoffs. Create tools to provide qualitative safety design information to engineers and decision makers. De-emphasize probabilistic methods, which are not accurate for complex systems.

- Create techniques and approaches that emphasize building safety into a system from the very beginning and provide better systems analysis and design tools that use these approaches. Decrease the current emphasis on adding protection to a completed design and trying to assure safety after-the-fact, i.e., after the design is complete: Neither is possible in complex systems.
- Create and use true sociotechnical approaches. Such approaches will require more integration of human factors and hardware/software engineering. It will also require better communication and ways for human factors engineers and hardware/software engineers to work together. Using STAMP as a basis, new modeling and analysis tools have been developed that provide a common model and communication process for all the engineers on a complex design project.
- Develop top-down, holistic approaches that allow us to handle the complexity of today's high-tech, complex systems. Closely related is to make sure that software and new technology can be handled by all our engineering and safety approaches.
- Improve our ability to create effective safety management systems and organizational and industry safety cultures.
- Create improved techniques for dealing with safety during operations and for using the massive amounts of data that can now be collected.
- Create new approaches to certification of safety for highly automated systems.
- Provide better education about safety for everyone. The fact that so many people have accepted Prof. Hollnagel's descriptions of safety engineering as Safety-I is an indication that both engineers and social scientists need a better education in the basic history and practices of safety engineering. Few classes are taught in universities about safety engineering beyond a course here and there on fault tree analysis and probabilistic risk analysis. Many academics don't know much about safety engineering themselves and thus cannot incorporate it into their university classes. This needs to be fixed.

References

Air Force Space Division, *System Safety Handbook for the Acquisition Manager*, SDP127-1, January 12, 1987.

Amalberti, R. The paradoxes of almost totally safe transportation systems, *Safety Science* 37:109-126, 2001.

Benner, Ludwig. Accident perceptions: Their implications for accident investigations. In Ted S. Ferry (editor) Readings in *Accident Investigation: Examples of the Scope, Depth, and Source*, pages 3-6. Charles C. Thomas Publisher, Springfield, Ill, 1984.

Calder, J. Scientific accident prevention, *American Labor Legislative Review*, 1:14-24, January 1911.

Checkland, Peter. *System Thinking, Systems Practice*. John Wiley & Sons, New York, 1981.

Cooper, J.H., Accident-prevention devices applied to machines, Transactions of the ASME, 12:249-264, 1891.

Drucker, Peter. "They're Not Employees, They're People," *Harvard Business Review*, 80(2):70-77, 2002.

Ferry, Ted S. *Safety Program Administration for Engineers and Managers*, Charles C. Thomas Publisher, Springfield, Ill, 1984.

Gloss, David S. and Miriam Gayle Wardle. *Introduction to Safety Engineering*, John Wiley & Sons, New York, 1984.

Gordon, John. The epidemiology of accidents. A presentation to the American Public Health Association Annual Meeting , Boston, Mass., Nov. 12, 1948 and later published in the *American Journal of Public Health*, August 29, 2011.

Griffin, Michael. *System Engineering and the Two Cultures of Engineering*, Boeing Lecture, Purdue University, 18 March 2007.

Hammer, Willie. *Product Safety: Management and Engineering*. Prentice-Hall Inc., Englewood Cliffs, M.J. 1980.

Hansen, Carl M. *University Safety Standards*, Universal Safety Standards Publishing Company, New York, 1914.

Hansen, Carl M. Standardization of safeguards, in *Proceedings Fourth Safety Congress*, pages 139-146, 1915.

Heinrich, H.W. *Industrial Accident Prevention: A Scientific Approach*, McGraw-Hill, New York, 1931.

Hollnagel, Erik. *Safety-I and Safety-II*, CRC Press, 2014

Hollnagel, Erik. *Safety-I I in Practice*, Routledge, 2018

Hope, S. et.al. Methodologies for hazard analysis and risk assessment in the petroleum refining and storage industry. *Hazard Prevention*, pages 24-32, July/August, 1983.

Lederer, Jerome. How far have we come? A look back at the leading edge of system safety eighteen years ago, *Hazard Prevention*, pp. 8-10, May/June 1986.

Leveson, Nancy, *Safeware*, Addison-Wesley, 1995.

Leveson, Nancy, *Engineering a Safer World*, MIT Press, 2012.

Nancy G. Leveson and John P. Thomas*, STPA Handbook*, 2018, http://psas.scripts.mit.edu/home/materials

Nancy Leveson, Nicolas Dulac, David Zipkin, Joel Cutcher-Gershenfeld, John Carroll, and Betty Barrett, "Engineering resilience into safety critical systems" in Erik Hollnagel, David Woods, and Nancy Leveson (eds.), *Resilience Engineering*, Ashgate Publishing, 2006.

Lewycky, Peter. Notes toward an understanding of accident causes. *Hazard Prevention*, pages 6-8, March/April 1987.

Manuele, Fred A. Accident investigation and analysis. In Ted S. Ferry (editor) Readings in *Accident Investigatioon: Examples of the Scope, Depth, and Source*, pages 201-211. Charles C. Thomas Publisher, Springfield, Ill, 1984.

Mill, John Stuart. *A System of Logic, Ratiocinative, and Inductive: Being a Connective View of the Principle of Evidence, and Methods of Scientific Inquiry*. J.W. Parker, London, 1843.

Miller, C.O. A comparison of military and civilian approaches to aviation safety, *Hazard Prevention*, May/June: 29-34, 1985.

Miller, C.O. The broader lesson from the Challenger. *Hazard Prevention*, pp. 5-7, January/February 1987.

Mulhouse Society, Mulhouse, Alsace-Lorraine. *Collection of Appliances and Apparatus for the Prevention of Accidents in Factories*, Society for the Prevention of Accidents in Factories, 1895.

National Nuclear Energy Policy Group, *Nuclear Power Issues and Choices*, Ballinger, Cambridge, MA, 1977

Petersen, Dan. *Techniques of Safety Management*, McGraw-Hill Book Company, New York, 1971

Petroski, Henry, *To Engineer is Human: The Role of Failure in Successful Design*, Vintage Books, New York, 1992

Petroski, Henry, *Success Through Failure: The Paradox of Design*, Princeton University Press, Princeton, New Jersey, 2006.

Rasmussen, Jens. Approaches to the control of the effects of human error on chemical plant safety. In *International Symposium on Preventing Major Chemical Accidents*, American Inst. of Chemical Engineers, February 1987a.

Rasmussen, Jens. The definition of human error and a taxonomy for technical system design. In Jens

Rasmussen, Keith Duncan, and Jacques Laplat (eds.), New Technology and Human Error, pages 23-30, John Wiley & Sons, New York, 1987b.

Rasmussen, Jens. Human error and the problem of causality in analysis of accidents, in D.E. Broadbent, J. Reason, and A. Baddeley, *Human Factors in Hazardous Situations*, pp. 1-12, Clarendon Press, Oxford, 1990.

Rasmussen, Jens. Risk management in a dynamic society: A modelling problem. *Safety Science* 27(2/3):183-213, 1997

Roberts, Verne L. Defensive design, *Mechanical Engineering*, pp. 88-93, September 1984.

Rogers, William P. *Introduction to System Safety Engineering*, John Wiley & Sons, New York, 1971.

Rouhiainen, Veikko. *The Quality Assessment of Safety Analysis*, Technical Report Publications 61, Technical Research Center of Finland, Espoo, Finland, 1990.

Stieglitz, William I. Engineering for safety, *Aeronautical Engineering Review*, February 1948.

Suokas, Juoko. Evaluation of the quality and safety and risk analysis in the chemical industry, *Risk Analysis*, 8(4):581-591, 1988.

Thomson, J.R. *Engineering Safety Assessment: An Introduction*. John Wiley & Sons, New York, 1987.

Thygerson, Alton L. *Accidents and Disasters: Causes and Countermeasures*, Prentice-Hall, Englewood Cliffs, N.J. 1977.

Von Bertalanffy, Ludwig. *General Systems Theory: Foundations*. Braziller, New York, 1969.

Weinberg, Alvin M. Science and trans-science, *Minerva*, 10:209—222, 1972.

Weinberg, Gerald. *An Introduction to General Systems Thinking*, John Wiley & Sons, New York, 1975.

Weiner, Norbert, *Cybernetics: or the Control and Communication in the Animal and in Engineering*, MIT Press, Cambridge, MA, 1965.

Appendix A: System Theory vs. Complexity Theory

Systems theory is a set of principles that can be used to understand the behavior of systems (as defined above), whether they be natural (for example, biological) or man-made systems. *Systems thinking* is the term often used to describe what people are doing when they apply systems theory principles. Systems theory is explained in more detail in the main part of the paper in the discussion of accident causality models.

Complexity theory grew out of systems theory and other concepts in the 1960s and is usually associated with the Santa Fe Institute and researchers working there. Some commonalities exist between Systems Theory and Complexity Theory as the latter was based on the former. Both include terms like emergence and focus on complex system behavior as a whole rather than on reduction or decomposition into components. The basic components of System Theory are emergence, hierarchy, communication and control and these also are included in Complexity Theory. Both Systems Theory and Complexity Theory also include concepts of feedback and feedforward control, adaptability, nonlinear interactions, and constraints. Both reject reductionism or decomposition as a principle for understanding system behavior.

There are, however, significant differences. Complexity Theory was created to describe natural systems where seemingly independent agents spontaneously order and reorder themselves into a coherent system using laws of nature that we do not yet fully understand. Systems theory, in contrast, is more appropriate for man-made and designed systems where the system is purposely created by humans using some engineering or design process and where the basic design is known and changes are controlled. Systems theory is also widely used in biology (where the "systems" are not engineered or designed), and, in fact, one of the founders of Systems Theory (Ludwig von Bertalanffy) was a biologist. Engineered systems, such as an aircraft or a car, do not spontaneously reorder themselves. Neither do most biological and even natural systems, although they may appear to do so because we do not completely understand how they work.

Another important difference is that Systems Theory considers all systems as displaying emergent behavior while Complexity Theory divides systems into four types: simple, complicated, complex, and chaotic, each with different degrees or types of emergence. Only complex and chaotic systems in complexity theory display emergent behavior. This distinction does not apply to engineered systems. In Systems Theory, all systems can and usually do display emergent behavior.

As such, Systems Theory appears to be most appropriate for engineered or designed systems (including organizations and designed social systems) while Complexity Theory is most appropriate for natural systems where the design is unknown, such as the weather, and for some sociological systems, such as communities, that are not designed and where there is a lack of order and it is very hard or impossible to predict the emergent behavior. Many (perhaps most?) social systems, such as regulation of aviation or the management structure of a company, however, are designed.

Complexity Theory deals with behavior that cannot be designed but instead must be experienced and studied as it arises. Engineered systems *are* designed, the design is known, and it is usually possible to predict the emergent behavior. I believe that Complexity Theory frameworks provide a poor basis for the goals related to improving safety and other emergent system properties in engineered sociotechnical systems.