Improving the Standard Risk Matrix using STPA

by Nancy G. Leveson
Cambridge, Massachusetts

This paper discusses the limitations of the standard risk matrix, and suggests some changes to the risk matrix and its use to improve the accuracy of the results.

What is the Risk Matrix and How is it Used?

A risk matrix is commonly used to define the level of risk for a system or specific events, and to determine whether the risk is sufficiently controlled. The matrix almost always has two categories for assessment: severity and likelihood (or probability). Figure 1 shows an example; there are many variants, but most are similar to the example shown.

While some potential problems occur in defining severity, the biggest problems arise in trying to assess likelihood, which is impossible to predict with any accuracy. While likelihood might be defined using historical events, most systems today differ significantly from the same systems in the past; for example, with much more extensive use of software or the use of new technology and designs. In fact, the usual reason for creating a new system is that existing systems are no longer acceptable. Historical data tells us only about the past, but the risk matrix is usually used to predict the future.

Even if the system's design does not change in the future, the way the system is used or the environment in which it is used will almost always change over time. Systems migrate toward higher risk over time for a variety of reasons [Ref. 1]. The past is a poor predictor of the future — and estimating future changes, along with their impacts, is essentially impossible.

How Accurate are Risk Matrix Results?

While standard Probabilistic Risk Analysis (PRA) has been subjected to scientific evaluation a few times — with poor results each time [Refs. 2 & 3] — we are unaware of any scientific evaluation of the accuracy, reliability and predictive capability of the risk matrix itself. Evidence of accuracy may be drawn from practical use of the risk matrix or from general technical limitations identified by experts. Each of these is discussed here.

Practical Limitations in the Use of Risk Matrices

We have anecdotal evidence that we have collected ourselves on real defense projects [Refs. 4 & 5] and in other experiences of using risk matrices in industry. We accumulated our experiences in applying systems theoretic process analysis (STPA) to real systems and then compared the results with the official risk assessment in the safety assessment report (SAR). The examples in this section stem from our experimental application of

RISK ASSESSMENT MATRIX

SEVERITY PROBABILITY	Catastrophic (1)	Critical (2)	Marginal (3)	Negligible (4)
Frequent (A)	**HIGH**	**HIGH**	**SERIOUS**	**MEDIUM**
Probable (B)	**HIGH**	**HIGH**	**SERIOUS**	**MEDIUM**
Occasional (C)	**HIGH**	**SERIOUS**	**MEDIUM**	**LOW**
Remote (D)	**SERIOUS**	**MEDIUM**	**MEDIUM**	**LOW**
Improbable (E)	**MEDIUM**	**MEDIUM**	**MEDIUM**	**LOW**
Eliminated (F)	Eliminated			

Figure 1 — A Standard Risk Matrix from MIL-STD-882E.

STPA to the Black Hawk helicopter (UH-60MU) and a naval vessel.

One common problem is that often the events assessed are only component failures, e.g., loss of external communication or breaking piston nuts, versus more general system hazards such as aircraft instability or inadequate separation from terrain. In the risk assessment for the Black Hawk, for example, a failure analyzed was "loss of displayed flight state information" [Ref. 6], rather than the hazards that this loss might lead to such as unsafe control actions provided by the flight crew or loss of control. And what about non-failures where the system components satisfied their requirements but hazards arose from interactions among the system components?

“There are other serious practical problems in the estimation of severity and likelihood of failures. One common complication is that assumptions may be made that operators, such as the flight crew, will not only recognize the failure (or hazard) but will also respond appropriately. Ironically, accidents often are blamed on inadequate flight crew or operator behavior while at the same time, the assumption that they will behave correctly is made in the risk assessment. Clearly, there are many cases where this assumption will not hold.”

Another problem with considering only failures rather than hazards is that *individual* failures are usually considered, but combinations of low-ranked failures are not. For example, consider a situation where a degraded visual environment occurs, along with a loss of altitude information, heading indication, airspeed indication, aircraft health information or internal communication. Individually, each of these losses may not result in an accident, particularly if it is assumed (as is often the case) that the pilots will react appropriately. When multiple losses occur simultaneously, however, the likelihood of an accident may be significant. Looking at each loss separately in the risk matrix can lead to a low system risk assessment due to a low probability of occurrence and low severity level of each of the individual (single-point) failures. There is also usually an assumption of independence of the failures and often a lack of consideration of common failure modes. It is not surprising that such combination failures are not considered, given the large number of failures possible in any realistic system; assessing all combinations becomes prohibitively expensive and usually infeasible. However, not considering combinations of failures affects the accuracy of the results.

There are other serious practical problems in the estimation of severity and likelihood of failures. One common complication is that assumptions may be made that operators, such as the flight crew, will not only recognize the failure (or hazard) but will also respond appropriately. Ironically, accidents often are blamed on inadequate flight crew or operator behavior while, at the same time, the assumption that they will behave correctly is made in the risk assessment. Clearly, there are many cases where this assumption will not hold. The mental model of the system operator (a general component of *situation awareness*) plays an important role in accidents. In aircraft, for example, the flight crew must receive, process and act on numerous sources of feedback about the state of the aircraft in order to interact correctly and safely with the various vehicle and mission systems. Time to perform this decision making may be limited. The interaction of control mode displays, pedal and other control positions, reference settings for various operating modes, and other visual and proprioceptive feedback can lead to flight crew mode confusion and an accident — particularly when external visual feedback is degraded. Omitting these interactions and assuming that the crew will (and can) always do the correct thing can lead to inaccurate risk assessments.

But problems exist not only in unrealistic assumptions about human behavior. Similar unrealistic assumptions often exist for hardware and software. As an example, in the official risk assessment for the Black Hawk, the failure "loss of displayed flight state information" was identified as catastrophic in severity, but improbable in likelihood. The only mitigations considered were hardware redundancy and a high level of rigor in the software development. Note, however, that redundancy does not prevent hardware design errors — only random "wear-out" failures. In addition, software is pure design and thus does not "wear out," so redundancy is not useful for software.

What about "rigor of development," which is assumed, often incorrectly, to increase safety? Almost all accidents involving software stem from flawed requirements often involving omissions, and not from flawed software implementation or assurance practices. The level of rigor in software development will have no impact on the completeness and accuracy of the software requirements — these are system engineering responsibilities. One of the reasons most software-related accidents arise from flawed requirements is that developing software requirements is a difficult and potentially flawed process. Rigor of software development will not help here.

The official Black Hawk risk assessment used these assumptions to identify as "relatively low likelihood" a loss of attitude information, loss of heading indication, loss of aircraft health information, loss of external communications and loss of internal communications. Note, however, that some of these losses have been implicated in Black Hawk accidents. As an example, the 1994 friendly-fire accident involved a loss of communication between the Black Hawk crew, AWACS controllers and the F-15 pilots involved. This set of conditions was not included in the official Black Hawk risk matrix, but was included in the STPA hazard analysis because the STPA analysis examined non-failure scenarios and did not assume perfect behavior on the part of the flight crews.

Events may appear improbable only if some of the likely factors involved — such as software requirements flaws and aspects of human behavior — are not considered. The Black Hawk STPA analysis found many non-failure scenarios (in addition to the previous example) that can lead to a hazardous system state but were not considered at all in the official risk assessment. It also identified realistic scenarios where the flight crew would not behave appropriately and suggested additional controls to prevent unsafe behavior, as well as important safety requirements for the software. Finally, and perhaps most disturbing, STPA identified realistic and relatively likely scenarios leading to all the specific failures dismissed as improbable in the official risk assessment. The omission of these types of scenarios will lead to an inaccurate risk assessment.

Similar limitations in the official risk assessment were identified in the software-intensive positioning system for a new naval vessel [Ref. 5]. Additional risk assessment limitations, however, existed in this system. For example, the likelihood of a loss can differ significantly depending on the external environment in which a failure occurs. But that factor is not usually considered in the risk matrix. In addition, likelihood and severity may be so entangled (for example, through the external environment) that again they cannot be evaluated along separate and independent dimensions. Using the results of the official risk assessment and ignoring the STPA analysis, this naval vessel was put into operation. Within two months, it collided with a nuclear submarine, producing extensive damage. The scenario that accounted for the accident sounds like one that was identified by STPA but ignored — along with the entire STPA analysis.

Technical Limitations

The rather dismal accuracy in the use of the current risk matrix stems from technical limitations. Space limitations prevent further details about the mathematical and other limitations, but they can be summarized as follows:

- A lack of granularity in the risk matrix makes it suited only for ranking events rather than providing information needed to make decisions about controlling the risk for specific events.
- The two ordinal scales make it impossible to do sophisticated calculations with the entries. The risk matrix can indicate only the category in which an event fails.
- Events that are potentially catastrophic but have a low estimated frequency tend to fall off the scale and get less attention than they deserve, particularly given the inaccuracy of most likelihood estimates.
- As the past is a poor estimate of the future, particularly because the way systems are used and the environment in which they are used will change over time, accurate prediction about operational behavior is not possible using a risk matrix.
- Poor resolution results from qualitative categories that are ill-defined and subjective, and can lead to assigning identical ratings to quantitatively different events.
- For risks with negatively correlated frequencies and severities, risk matrices can be "worse than useless," leading to worse-than-random decisions [Ref. 7].
- Categorizations of severity cannot be made objectively for uncertain consequences. In these cases, a worst-case analysis leads to high severity for every event. At the same time, expected case evaluation may be optimistic.
- The subjective interpretations of the categorizations of severity and likelihood (particularly likelihood) can lead to different categorizing of the events by different users.
- Risk matrices produce arbitrary risk rankings when they depend on the design of the matrix itself, such as how large the bins are and whether one uses an increasing or decreasing scale. Changing

> “As a final example, a search for possible causes is often stopped once one possible cause or explanation for an event has been identified. If that first possible cause is not compelling, stopping the search at that point leads to nonidentification or underestimation of risk of other more plausible and compelling causes.”

the scale can change the answer. The errors in expert predictions are exacerbated by the additional errors introduced by the scales and matrices.

- Likelihood can, and often does, ignore or discount certain types of causal factors, such as operator errors, management decisions and, sometimes, software behavior. Random failures of hardware are usually over-emphasized.

Some of the most interesting limitations stem from what Kahneman and Tversky call heuristic biases [Refs. 8 & 9]. Kahneman and Tversky are psychologists who studied how people actually do risk evaluation. It turns out that humans are really terrible at estimating risk, particularly likelihood. For example, people tend to deny uncertainty and vulnerability and over-rate estimates that conform to their previous experience or views (called *confirmation bias*). As another example, people often will construct their own simple causal scenarios of how the event could occur, using the difficulty of producing reasons for an event’s occurrence as an indicator of the event’s likelihood. If no plausible cause or scenario comes to mind easily, an assumption may be made that the event is impossible or highly unlikely.

People also tend to identify simple, dramatic events rather than causes that are chronic or cumulative. Dramatic changes are given a relatively high probability or likelihood, whereas a change resulting from a slow shift in social attitudes, for example, is more difficult to imagine and thus is given a lower likelihood. As a final example, a search for possible causes is often stopped once one possible cause or explanation for an event has been identified. If that first possible cause is not compelling, stopping the search at that point leads to non-identification or underestimation of risk of other more plausible and compelling causes.

One way to overcome these biases is to provide those responsible for creating the matrix with better information about the scenarios that can lead to the loss event, perhaps through a structured process like STPA to generate the scenarios. Another is to change the risk matrix itself to reflect a more general and practical definition of risk. Both of these potential ways forward are discussed in the next section.

Potential Improvements

There are two possible ways to improve the standard risk matrix while making the fewest changes to what is done today: 1) use hazards instead of failures and 2) use better information about potential causal scenarios to improve severity and likelihood estimates.

Use Hazards Rather than Failures

Some of the inaccuracy in risk matrix severity evaluations stems from the fact that the relationship between individual failures and accidents (losses) may not be obvious and may require a lot of work to determine. Assigning severity and likelihood to hazards, rather than to failures, provides a more direct path to the ultimate goal of the risk matrix, which is to assess risk of losses, rather than component or even system unreliability. Component or system reliability is not equivalent to system safety, although there are overlaps. In many cases, system reliability can conflict with system safety; i.e., increasing one may decrease the other.

Traditionally, in system safety engineering, safety is defined in terms of hazards, not failures. Prioritization of hazard severity starts with the assessed severity of the loss (accident) by stakeholders — hazards are then associated with the prioritized losses. This process is easier and more straightforward than starting with attempting to prioritize the severity of system or component failures by tracing them to accidents. There are usually an enormous number of potential failures in a complex system, and the consequences are not always clear. Of course, hazards that result from design errors or other aspects of the system that do not involve failures will be omitted from consideration.

As an example of the latter, consider the helicopter de-ice function. The final SAR [Ref. 6] on a Black Hawk upgrade included a failure of the aircraft's Auxiliary Power Unit (APU) resulting from APU chaffing. This failure is important because the APU is used when the loss of one generator occurs during blade de-ice operations. While APU chaffing *can* prevent the de-ice function from operating, there is another scenario — found using STPA — that could prevent the blade de-ice function when the APU has not failed. Consider the following unsafe control action (UCA):

UCA: The flight crew does not switch the APU (Auxiliary Power Unit) generator power ON when either GEN1 or GEN2[1] *are not supplying power to the helicopter and the blade de-ice system is required to prevent icing.*

There are several causal scenarios and factors that could lead to this unsafe control beyond APU chaffing or even component failure [Ref. 4]. These are not included in the official Black Hawk SAR, but they need to be factored into any risk assessment and used to develop design, testing and operational requirements. The new scenarios for this UCA could lead to requiring the software and hardware designers to assign higher criticality to hardware and software that is used to generate and display specific cautions to the crew, and to improve the design of the role the flight crew plays during operations. Considering only failures as the cause of hazards and accidents severely distorts the risk assessment, and the results are likely to be inaccurate for today's increasingly complex systems.

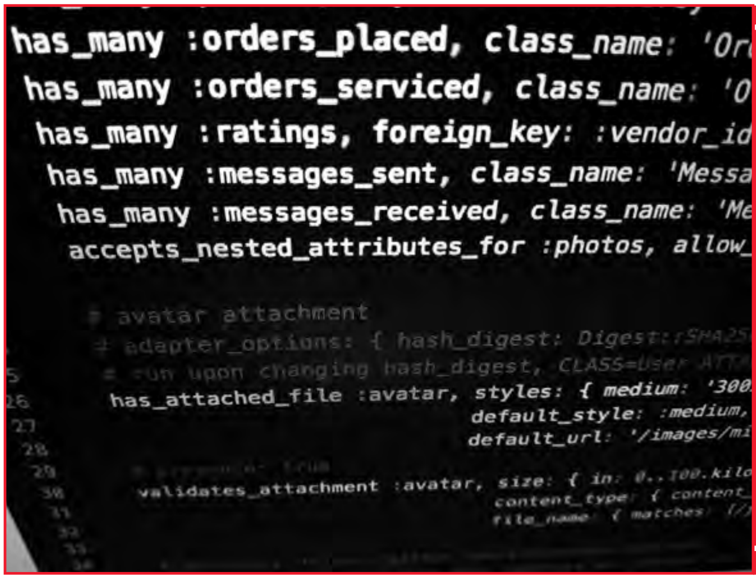
The change being suggested here, then, is to start from a prioritized list of stakeholder-identified accidents or system losses. Then, the high-level system hazards (conditions or states) that can lead to these accidents are identified. This process is consistent with MIL-STD-882 (in all its incarnations), along with many other safety standards. The severity and likelihood of the hazards are then assessed. Only the failures that can lead to hazards (which can be identified by STPA) need be considered, not all failures. In addition, hazards resulting from causal scenarios, including non-failures (e.g., design errors), must be included in the assessment. These more general scenarios may be derived from STPA or other analysis methods that provide similar results.

Define Likelihood as Strength of Potential Controls

Starting from hazards makes the evaluation of severity straightforward, as the hazards can be directly linked to the stakeholder-prioritized list of accidents or losses. That leaves the evaluation of likelihood as the remaining obstacle to more accurate risk assessment using the standard risk matrix. The heuristic biases described earlier explain why people often do a poor job of assessing risk. The biases arise because informal processes, i.e., heuristics, are used to estimate risk, particularly likelihood. One way to overcome such biases is to require following a structured process to identify scenarios and not allow stopping before full consideration of these scenarios in the risk assessment. Of course, one cannot ensure completeness in any non-mathematical process, but following a rigorous process, such as STPA, will result in reducing shortcuts and biases, along with fuller consideration of potential causal scenarios.

One problem in assessing likelihood is that little real design information is available at the beginning of the development process, when decisions about where to focus efforts are made. Without having the final detailed system design, it is not possible to determine the likelihood of an accident occurring. Even later, there are problems in assessing the likelihood of unsafe software or human behavior. One reason that component failures may be the focus of current risk assessment activities is that there is usually historical information about failures of standard components — although that does not guarantee that new designs will have the same failure likelihoods. Solving the wrong problem because we know the solution is like the old joke about a man who comes across a drunken individual crawling around on a sidewalk underneath a streetlight, looking for his lost wallet. The man offers to help and asks where the he lost his wallet, and he points to the other side of the street. When the man asks why he is looking in a place different from where he dropped the wallet, he explains that the light is better here. We need to get better risk assessments by focusing on the actual problem rather than a different one we know how to solve.

[1] Redundant APU generators



```
has_many :orders_placed, class_name: 'Or
 has_many :orders_serviced, class_name: 'O
  has_many :ratings, foreign_key: :vendor_id
   has_many :messages_sent, class_name: 'Messa
    has_many :messages_received, class_name: 'Me
     accepts_nested_attributes_for :photos, allow_

      # avatar attachment
      # adapter_options: { hash_digest: Digest::SHA25
      # run upon changing hash_digest, CLASS=User ATTA
      has_attached_file :avatar, styles: { medium: '300
                                 default_style: :medium,
                                 default_url: '/images/mi

         validates_attachment :avatar, size: { in: 0..100.kilo
                                   content_type: { content_
                                   file_name: { matches: [/
```

“The design of the software and hardware also must be included in the risk assessment. Current approaches to handling software, such as assigning levels of rigor to software development, have no technical or scientific basis, as mentioned earlier. Simply assuming that software-related risk is adequately reduced or eliminated by rigorous development is not realistic and does not reflect either research results or real engineering experience.”

Potentially, scenarios generated by STPA can provide better information with which to evaluate the likelihood of hazards occurring. What types of information will be created? Consider the following example from the Black Hawk STPA analysis. One unsafe control action (UCA) is that:

UCA: The Flight Crew does not deflect pedals sufficiently to counter torque from the main rotor, resulting in the Flight Crew losing control of the aircraft and coming into contact with an obstacle in the environment or the terrain.

One of the causal scenarios that could lead to this unsafe control action might be:

Scenario 1: The Flight Crew is unaware that the pedals have not been deflected sufficiently to counter the torque from the main rotor.

The Flight Crew could have this flawed process model because:

a) The flight instruments are malfunctioning and providing incorrect or insufficient feedback to the crew about the aircraft state during degraded visual conditions.
b) The flight instruments are operating as intended, but are providing insufficient feedback to the crew to apply the proper pedal inputs to control heading of the aircraft to avoid obstacles during degraded visual conditions.
c) The Flight Crew has an incorrect mental model of how the flight control systems (FCS) will execute their control inputs to control the aircraft and how the engine will respond to the environmental conditions.
d) The Flight Crew is confused about the current mode of the aircraft automation and is thus unaware of the actual control laws that are governing the aircraft at this time.
e) There is incorrect or insufficient control feedback.

Each of these causal factors can be used to create requirements and design features to reduce their likelihood and thus the likelihood of the UCA and the hazard. The key impact on risk assessment is that likelihood can then be based on the *strength of the potential controls.* In Scenario 1, factor (a) could be controlled through redundancy and fault-tolerant design. Factor (b) could be controlled by interface design (as evaluated by a human factors expert). Factor (c) will be impacted by interface design and also by training. Factors (d) and (e) can be controlled through system design (both hardware and software and their interactions) and through design of feedback. However, a way to link these factors to likelihood is needed. A few are suggested in the next section.

The example shown so far focuses on the interaction of the flight crew and the aircraft controls. The design of the software and hardware also must be included in the risk assessment. Current approaches to handling software, such as assigning levels of rigor to software development, have no technical or scientific basis, as mentioned earlier. Simply assuming that software-related risk is adequately reduced or eliminated by rigorous development is not realistic and does not reflect either research results or real engineering experience. Using the approach to risk assessment described here, software-related risk assessment can be handled in the same way as hardware- and human-related risk assessment.

As an example, consider the following UCA identified by STPA for the Black Hawk:

UCA: One or more of the FCCs (flight control computers) command collective input to the hydraulic servos too long, resulting in an undesirable rotor RPM condition and potentially leading to the hazard of violating minimum separation from terrain or the hazard of losing control of the aircraft.

There are at least five causal scenarios that could lead to this unsafe control action:

Scenario 1: The FCCs are unaware that the desired state has been achieved and continue to supply collective input. The FCCs could have this flawed process model because:

a) The FCCs are not receiving accurate position feedback from the main rotor servos.
b) The FCCs are not receiving input from the ICUs to stop supplying swashplate input.

Scenario 2: The FCCs do not send the appropriate response to the aircraft for particular control inputs. This could happen if:

a) The control logic does not follow intuitive guidelines that have been implemented in earlier aircraft, perhaps because requirements to do so were not included in the software requirements specification.
b) The hardware on which the FCCs are implemented has failed or is operating in a degraded state.

Scenario 3: The FCCs do not provide feedback to the pilots to stop commanding collective increase when needed because the FADEC (full authority digital engine control) is supplying incorrect cues to the FCCs regarding engine conditions.

Scenario 4: The FCCs do not provide feedback to the pilots to stop commanding collective increase when needed because the FCCs are receiving inaccurate NR (rotor rpm) sensor information from the main rotor.

Scenario 5: The FCCs provide incorrect tactile cueing to the inceptor control units (ICU) to properly place the collective to prevent low rotor RPM conditions.

While typically these STPA-generated scenarios would be used to identify appropriate FCC requirements and design constraints, the information could also feed into a risk assessment. For example, three safety requirements could be identified related to Scenario 1:

1. The FCCs must perform median testing to determine if feedback received from the main rotor servos is inaccurate.
2. The PR SVO FAULT caution must be presented to the Flight Crew if the FCCs lose communication with a main rotor servo.
3. The EICAS (engine-indicating and crew-alerting system) must alert the Flight Crew if the FCCs do not get input from the ICU every x seconds.

Risk of the hazard related to the UCA will be reduced by implementing these requirements and increased if they or other controls to reduce the occurrence of the UCA are not included in the design. The risk assessment then can use the strength of potential controls. At the simplest level, this assessment might involve differentiating between controls that eliminate the hazard versus those that try to detect and mitigate it.

Translating Strength of Controls into Likelihood

The problem of associating likelihood with strength of potential controls remains. In system concept development and in early decisions about the development process (e.g., where to invest resources), an estimate of the potential strength of designed controls for the scenarios generated by STPA would be used to assess likelihood. As the basic design decisions are made, testing is performed and the STPA analysis is refined; thus, the likelihood evaluations can be improved. In the end, the risk associated with the system during operations may be evaluated with much better accuracy than is currently possible.

Various strategies might be used to rank the strength of potential controls. One possible strategy (where 1 is the highest) is:

1. The causal factor can be *eliminated* through design and high assurance.
2. The occurrence of the causal factor can be *reduced or controlled* through system design
3. The causal factor can be *detected and mitigated* if it does occur through system design or through operational procedures
4. The only potential controls involve *training and procedures.*

This example ranking system may be too simple. A more sophisticated procedure might involve estimates of how well the causal factor has been handled within each of the four categories — for example, how thoroughly may the causal factor be mitigated. This procedure may improve the results better than simply assigning a single potential number (e.g., 1 - 4) for each category. For identified critical hardware failures, the potential impact of redundancy or other failure reduction or handling techniques on likelihood can be computed mathematically. But these are a subset of all the causal factors that STPA can identify. Other types of safety-enhancing techniques may not be so easily evaluated and may require "engineering judgment."

"A second example is a scheme devised for evaluating risk in a human-intensive NASA project involving air traffic control (ATC) enhancements. This case was almost the exact opposite for the manned space mission design in that the system engineering problem was not to create a new or safer system but to maintain the already high level of safety built into the current system. The goal was essentially not to degrade the safety of the current system when changes were made to it. The risk analysis, then, is aimed at evaluating the risk that safety will be degraded by the proposed changes and new automated tools."

In addition, combinations of the four types of control listed here might be used in likelihood estimates; e.g., design features included to reduce or control the factor, as well as operator training and procedures as a back-up should the hazard still occur. A combination of controls might lead to reduction of the assessed likelihood. Other ranking strategies or mappings to levels of risk are also possible.

There is an assumption here, of course, that these control strategies will impact the likelihood of the hazard or UCA occurring. But this assumption is better than the assumption that historical hardware failure rates will apply to the future (no matter the changes in the system itself or to the environment during operations), combined with either 1) omitting all the factors that do not involve hardware component failures or 2) making up probabilities for these factors out of thin air. It is also better than assuming that general "rigor of development" will eliminate risk.

Specialized risk assessment processes that are appropriate for specific types of systems can be developed. Chapter 10 of Engineering a Safer World (pages 321 to 327) describes the two such special approaches we have devised for past projects [Ref. 10]. The first was for a NASA contract to create and analyze architectural trade-offs for future manned space exploration missions. The system engineers wanted to include a safety assessment of potential architectures along with the usual factors, such as mass, that are used in evaluating candidate architectures. Little information was available at this early stage of system engineering, and, of course, historical information about past space exploration efforts was not useful because all the potential architectures involved new technology and new missions, which invalidated past experience and even created new hazards, such as the use of nuclear energy to power the spacecraft and surface rovers.

In the process devised to assess risk for this architectural trade study, hazards specific to each mission phase (e.g., launch or landing) were identified, along with some general hazards such as fire, explosion or loss of life support that spanned all or most of the mission stages. Once the hazards were identified, the worst-case loss associated with the hazard was evaluated for its impact on three categories: humans, mission and equipment. Environment, including damage to the Earth and planet surface environments, was originally included, but then eliminated, when project managers decided all the missions must comply with NASA's planetary protection standards and could not be part of a trade-off analysis. Other projects may want to include environmental impact in the risk analysis.

A severity scale was created for each of the three categories. As usual, severity is easier to handle than likelihood. In this case, the architectures and missions would involve things that had never been attempted and historical data was not relevant. Instead, mitigation potential was substituted for likelihood, as in the earlier example but in a more sophisticated way. Mitigatibility was evaluated by domain experts under the guidance of safety experts. Both the cost and difficulty of the potential mitigation strategy (in qualitative terms of low, medium and high) and its potential effectiveness (on a comparative scale from 1 to 4) were evaluated. Because hundreds of feasible architectures were generated by the system engineers, the evaluation process was automated, and weighted averages were used to combine mitigation factors and severity factors to come up with a final Overall Residual Safety-Risk Metric. This metric was then used to evaluate and rank the potential manned space exploration architectures. A detailed example can be found in *Engineering a Safer World* [Ref. 10].

A second example is a scheme devised for evaluating risk in a human-intensive NASA project involving air traffic control (ATC) enhancements. This case was almost the exact opposite for the manned space mission design in that the system engineering problem was not to create a new or safer system but to maintain the already high level of safety built into the current system. The goal was essentially not to degrade the safety of the current system when changes were made to it. The risk analysis, then, is aimed at evaluating the risk that safety will be *degraded* by the proposed changes and new automated tools. In this case, we created a set of criteria to rank various high-level architectural design features of the proposed set of new ATC tools on a variety of factors related to system risk. Again, the ranking was qualitative, and most criteria were ranked as high, medium, or low impact on the potential for a degradation of safety from the current high level.

Many of the criteria chosen involved human-automation interaction because of the nature of the application and the fact that new features being proposed primarily involved new automation to assist air traffic controllers. Example criteria included:

- **Safety margins.** Does the new feature have the potential for 1) an insignificant or no change in the existing safety margins, 2) a minor change, or 3) a significant change?
- **Situation awareness.** What is the potential for reducing situation awareness?
- *Skills currently used and those necessary to backup and monitor the new decision-support tools.* Is there an insignificant change (or no change), a minor change or a significant change in the controller skills?
- **Introduction of new failure modes and hazard causes.** Do the new tools have the same function and failure modes as the system components they are replacing? Are new failure modes and hazards introduced, but well understood so that effective mitigation measures can be designed? Or are the new failure modes and hazard causes difficult to control?
- **Effect of the new software functions on the current system hazard mitigation measures.** Can the new features render the current safety measures ineffective or are they unrelated to the current safety features?
- **Need for new system hazard mitigation measures.** Will the proposed changes require new hazard mitigation measures?

These criteria, and others, were converted into a numerical scheme so they could be combined and used in an early risk assessment of the changes being contemplated, along with their potential likelihood for introducing significant new risk into the system. The criteria were weighted to reflect their relative importance in the risk analysis.

For both of these specialized examples and others that might be devised, using STPA to identify causal scenarios will help provide better values for the criteria. By thinking through what risk means in your particular project, you can help identify better ways to evaluate it — particularly the likelihood component.

So far in this paper, and often in practice, focus is primarily on the risk involved in the engineered system design at system deployment. Risk will be affected by many other factors during manufacturing and operations, including manufacturing controls; designed maintainability and the occurrence of maintenance errors; training programs; changes over time in the environment in which the system is used; and consistency and rigor of management and of oversight by those tasked to oversee the operation of the system, etc. The risk of deployed systems is based on the system designers' assumptions about the operational environment. How realistic and accurate those assumptions are, how well those assumptions are communicated to users, and how rigorously the operational assumptions are enforced will have a large impact on system risk. Including the potential impact of these additional factors will result in improved initial risk assessments. In addition, tracking these factors can provide improved risk assessments over time if it is not possible to predict them perfectly during system development. The process of risk assessment need not stop when systems are deployed. Risk-based decisions are required throughout the system lifecycle. Castilho [Ref. 11] has devised what he calls Active STPA, which can be used during operations to identify leading indicators of changes that increase risk.

> “Risk will be affected by many other factors during manufacturing and operations, including manufacturing controls; designed maintainability and the occurrence of maintenance errors; training programs; changes over time in the environment in which the system is used; and consistency and rigor of management and of oversight by those tasked to oversee the operation of the system, etc.”

Conclusion

While the use of rigorously developed causal scenarios using STPA does not avoid all the problems with standard risk matrices, it does provide a more rational basis for categorizations. Fault trees and other hazard analysis techniques might be used here, but they typically cannot start until a detailed system design is available — which is late in the development process, when the use of the risk matrix to determine how to allocate development effort is not very helpful. In addition, adding risk reduction efforts late in development is expensive, extremely disruptive to project schedules, and usually less effective than if the controls are designed into the system from the beginning. STPA can be done earlier, at the point in concept development when the risk matrix is usually initially created and used.

A more important limitation is that fault trees and other hazard analysis techniques that assume accidents are caused by component failures leave out many (most?) of the causes of losses in today’s complex systems. The more comprehensive the causal scenarios that are used to assess likelihood, the better the estimates will be. ⬢

References

1. Rasmussen, Jens. “Risk Management in a Dynamic Society: A Modeling Problem,” *Safety Science*, Vol. 27, Issues 2-3, 183–213, 1997.
2. Lauridsen, K., I. Kozine, F. Markert, A. Amendola, M. Christou and M. Fiori. “Assessment of Uncertainties in Risk, 2002,” *Assessment of Uncertainties in Risk Analysis of Chemical Establishments*, Risø National Laboratory, Roskilde, Denmark, Risø-R-1344(EN), 2002.
3. Leveson, Nancy. *Safeware: System Safety and Computers*, Addison-Wesley, New York, 1995.
4. Abrecht, B., D. Arterburn, D. Horney, J. Schneider, B. Abel and N. Leveson. “A New Approach to Hazard Analysis for Rotorcraft,” AHS Technical Specialists’ Meeting on the Development, Affordability, and Qualification of Complex Systems, Huntsville Alabama, February 9 – 10, 2016.
5. Abrecht, Blake. *Systems Theoretic Process Analysis Applied to an Off-Shore Supply Vessel Dynamic Positioning System*, S.M. Thesis, Massachusetts Institute of Technology Dept. of Aeronautics and Astronautics Dept., 2016.
6. Sikorsky Aircraft Corporation. “Safety Assessment Report for the UH-60M Upgrade Aircraft, Document Number SER-703655,” January 3, 2012.
7. Cox, Anthony. “What’s Wrong with Risk Matrices,” *Risk Analysis*, Vol. 28, Issue 2, 497–512, 2008.
8. Kahneman D and A. Tversky. “On the Psychology of Prediction,” *Psychological Review*, Vol. 80, Issue 4, 237–51, 1973.
9. Kahneman, D., P. Slovic and A. Tversky. *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, New York, 1982.
10. Leveson, Nancy. *Engineering a Safer World*, MIT Press, Cambridge, Massachusetts, 2012.
11. Castilho, Diogo Silva. *A Systems-based Model and Processes for Integrated Safety Management Systems (I-SMS)*, Ph.D. Dissertation (in process), Massachusetts Institute of Technology Dept. of Aeronautics and Astronautics Dept., expected August, 2019.