Improving the Standard Risk Matrix: Part 1[1]

Prof. Nancy Leveson
Department of Aeronautics and Astronautics
MIT
leveson@mit.edu

Abstract: Part 1 of this White Paper describes the standard risk matrix and its limitations. It then suggests some changes to the risk matrix and its use in order to improve the accuracy of the results. Part 2 suggests larger changes in terms of the basic definition and evaluation of risk that may even more greatly enhance our ability to assess risk but also challenge our willingness to change.

What is the Risk Matrix and How is it Used?

A risk matrix is commonly used for risk assessment to define the level of risk for a system or specific events and to determine whether or not 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 in Figure 1.

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.

Figure 1 is derived from the fact that the standard definition of risk is severity combined with likelihood: risk = f(severity, likelihood).[2] In many ways, defining risk in terms of how it is quantified is unfortunate. It has hindered progress by limiting risk to a very narrow definition and disallowing alternatives and potential improvements by definition. This white paper suggests some alternative definitions and ways to assess risk. One simple example of an alternative is that risk is the lack of

[1] © Nancy G. Leveson, February 2019

[2]Sometimes risk is described as severity "multiplied" by likelihood, but of course multiplying two different types of measurements makes little sense mathematically.

certainty about an outcome, often the outcome of making a particular choice or taking a particular action. More about this later.

The classic risk matrix uses two ordinal rating scales: severity and likelihood. The problems arise in defining severity and likelihood. While risk is often thought of as a quantitative quality, in practice it is usually defined qualitatively, i.e., in terms of ordinal rating scales for severity and likelihood. Using qualitative scales can only give a qualitative scoring that indicates a category or box in which the event falls. This conception does not allow for sophisticated calculations or subtle differences.

Severity

Severity is usually defined as a set of categories such as:

Catastrophic: multiple deaths

Critical: one death or multiple severe injuries

Marginal: one severe injury or multiple minor injuries

Negligible: one minor injury

Of course, these categories are subjective and could potentially be defined in different ways by the stakeholders. For example, why is one death not catastrophic? What is a "severe injury"? Alternatively, or in addition, monetary losses may be associated with the severity categories, although that raises the moral and practical quandary of determining the monetary value of a human life.

Severity is relatively straightforward to define although there remains the problem of whether the worst-case outcome is considered, only "credible" outcomes, most likely outcome, or only predefined common events.

Using the worst case is the most inclusive approach, but concerns may be raised that it is too pessimistic and instead the worst *credible* outcome should be used. The latter raises the problem of how to define "credible" and can lead to a blurring of the distinction between severity and likelihood, making these two factors not truly independent in the assessment of risk.

A third approach is to use the most likely outcome, which again mixes severity and likelihood and reduces their independence. In many cases, people may not be aware that they are doing this and simply default to assigning severity according to what they thought were the most likely outcomes. In aircraft certification, SAE ARP 4761 has an example of a wheel brake failure on landing being assigned "no safety effect" if the brake failure is annunciated to the flight crew. An assumption is made that if the pilots know about the failure, they will be able to safely bring the aircraft to a stop by, perhaps, steering off the runway or taxiway onto grass. While this is the most likely outcome, it is easy to think of specific situations where the pilots will be unable to prevent an accident even if they know the brakes have failed.

A final possibility, considering only specific predetermined failures or events (called in the nuclear industry a *design basis event,* such as a pipe break or more generally a loss of coolant) can result in the risk assessment being highly optimistic and often unrealistic due to being too limited in what is considered.

Even more problematic is the common practice of assessing risk using failures rather than hazards. The severity of a single failure or even multiple ones may not be easily determined in a complex system. What is the severity of a "loss of heading information" or the severity of a "human error"? It depends on how the heading information is used and the conditions of other parts of the system and the environment at the time or the specific details of the human error and the conditions under which it is made. Using worst-case severity, nearly every failure can be argued to be potentially catastrophic although the opposite (underestimating severity of failures) seems to be much more common.

Another major problem is that "software failure" makes little technical sense. Software is a pure abstraction with no physical being. How does an abstraction fail? Even defining software failure as the case where the software does not satisfy its requirements, there are usually an enormous number of ways that software may not satisfy its requirements and therefore the severity of a "software failure" is impossible to determine or could reasonably be argued to potentially always lead to a catastrophic outcome in the worst case. Assessing risk in terms of hazards (discussed later) rather than failures overcomes some of these problems as hazards are by definition linked to specific types of accidents or losses, which the stakeholders can identify and prioritize.

Likelihood

More problems arise in defining *likelihood*. When the risk matrix is used for prediction, the goal is to estimate how often an event might happen in the future. That information is difficult or impossible to determine. While likelihood might be defined using historical events, most systems today differ significantly from the same systems in the past, for example, by 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 only tell us about the past but the risk matrix is usually used to predict the future. Just because something has not occurred yet does not provide an accurate prediction about the future, particularly when the system or its environment differs from the past. And most people do not believe that the likelihood of a software "failure" (defined in some way) can be determined before long use of the software. Given the experience we have had with software and other practical considerations, some would snidely suggest that the probability for a software "failure" is always "1." And software is in everything these days.

Even if the design itself 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. The concept of "migration toward higher risk over time" [Rasmussen 1997] argues against the applicability of the past as a determinant for the future. And estimating future changes along with their impacts is essentially impossible.

The example risk matrix in Figure 1 categorizes likelihood in terms of frequent, probable, occasional, remote, improbable, and eliminated (or impossible). These categories usually need to be defined more precisely, such as one common approach in military systems:

Frequent: likely to occur frequently

Probable: Will occur several times in the system's life

Occasional: Likely to occur sometime in the system's life

Remote: Unlikely to occur in system's life, but possible

Improbable: Extremely unlikely to occur

Impossible: Equal to a probability of zero

As the reader can easily see, these definitions are not terribly helpful and simply restate the problem in a different but equally vague form. This same criticism holds for most of the attempts to define qualitative likelihood categories.

Sometimes the qualitative categories are associated with probabilities. An example might be using the categories *1.0E-9 and higher, between 1.0E-6 to 1.0E-8, and 1.0E-5 and lower*. This probabilistic assignment, however, does not eliminate the question of whether the probabilities can be determined in advance (i.e., before long operational use of the system).

Use of the Risk Matrix

Once the categories are determined, using the risk matrix involves assigning various types of events to appropriate boxes and thus "assessing" their risk. The events usually involve failures although hazardous conditions or states may also be used.

If the system has not yet been designed or is in the process of being developed and tested, the risk matrix category for the different events may be used to determine the amount and type of effort to apply in order to prevent those events from occurring. It may also be used to evaluate the effort required with respect to standard design processes mandated by the customer (e.g., level of rigor in development). There are, of course, serious questions about whether general level of rigor actually results in measurable differences in risk. This commonly accepted conclusion has never been proven and, at least for software, is almost certainly incorrect.

At the end of development, risk assessment might be used to make decisions about whether the risk is sufficiently controlled and to make decisions about certification, deployment and operational use of the system.

How Accurate is the Standard Risk Matrix?

While standard Probabilistic Risk Analysis (PRA) has been subjected to scientific evaluation a few times, with very poor results each time [Lauridsen et.al. 2002, Leveson 1995], 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 in this section.

Practical Limitations in the Use of Risk Matrices

We have some anecdotal evidence that we have collected ourselves on real defense projects [Abrecht et.al. 2016, Abrecht 2016] and in other experiences we have in the use of risk matrices in industry. The goal is not to criticize the particular engineers involved—they were simply following today's standard practices. Instead, the goal is to point out the practical limitations of risk matrices as they are defined and used today. The author has found the same flaws in the hundreds of risk matrices she has seen during her long career in this field.

One common problem is that often the events assessed are only component failures, e.g., loss of external communication or breaking piston nuts, and not the more general system hazards such as aircraft instability or inadequate separation from terrain. In the risk assessment for the Black Hawk helicopter, for example, a failure analyzed was "loss of displayed flight state information" [Sikorsky 2012] 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?

Another problem with considering only failures rather than hazards is that *individual* failures are usually considered but not combinations of low-ranked failures. For example, consider the situation where a degraded visual environment occurs as well as 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 for 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, of course, 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 beyond those described so far that occur in estimation of severity and likelihood of failures. One common complication is that assumptions may be made that the flight crew will not only recognize the failure (or hazard) but will also respond appropriately. Ironically, accidents are often blamed on inadequate flight crew or operator behavior while at the same time assuming they will behave correctly 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 upon 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 very limited. The interaction of control mode displays, pedal and other control position, 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 very inaccurate risk assessments.

But the problems are not just 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 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"? 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 the 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 such 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.

It appears that events may only appear improbable 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 example above) 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 the 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 a very inaccurate risk assessment.

Similar limitations in the official risk assessment were identified in the software-intensive positioning system for a new naval vessel [Abrecht 2016]. 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 (as was the entire STPA analysis).

Technical Limitations:

The rather dismal accuracy in the use of the current risk matrix stems from technical limitations. The goal of this white paper is not to go into details about the mathematical and other limitations, but they can be summarized as follows:

- The lack of granularity in the risk matrix makes it only suited for ranking events rather than providing the 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 only indicate in which category an event fails.
- What happens to events that are potentially catastrophic but have a very low estimated frequency? They tend to fall off the scale and get less attention than they deserve, particularly given the inaccuracy of most likelihood estimates.
- As mentioned, 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. Therefore, 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 what are quantitatively very different events.
- For risks with negatively correlated frequencies and severities, risk matrices can be "worse than useless," leading to worse-than-random decisions [Cox 2008].
- 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 very optimistic.
- The subjective interpretations of the categorizations of severity and likelihood (particularly likelihood) can lead to very different categorizing of the events by different users.
- Critics have shown that risk matrices produce arbitrary risk rankings because 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 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 [Kahneman and Tversky 1973; Kahnemann, Slovic, and Tversky 1982.]. 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. Here are a few of the relevant heuristic biases that have been described:

- *Confirmation Bias*: People tend to pay more attention to information that supports their views than to evidence that conflicts with them. The result is that people tend to deny uncertainty and vulnerability and overrate estimates that conform to their previous experience or views.
- *Availability Heuristic*: People tend to base likelihood judgments of an event on the ease with this instances or occurrences of that or similar events can be brought to mind. While this heuristic may often be a reasonable one to use, it can also lead to systematic bias. For example, psychologists have found that judgments of risk of various hazards or events will tend to be correlated with how frequently they are mentioned in the news media.
- *Ease of scenario generation*: People will often 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.
- *Difficulty in predicting cumulative causes*: People 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.
- *Conjunction fallacy*: An outcome paired with a likely cause is often judged to be more probable than the outcome alone, even though this conclusion violates the laws of probability.
- *Incomplete search for possible causes*: A search is often stopped once one possible cause or explanation for an event has been identified. If that first possible cause is not very compelling, stopping the search at that point leads to nonidentification or underestimation of risk of other more plausible and compelling causes.
- *Defensive avoidance*: Rejection or downgrading of risk categorizations that conflict with other pressing goals. The author has sat in meetings and watched managers systematically re-bin risks into lower categories due to budget and schedule pressures that have little to do with the actual risk being evaluated. The desire for lower categorization of risk can outweigh and suppress objectivity.

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 Alternatives to the Standard Risk Matrix

As suggested, there are two possible ways to improve the standard risk matrix: (1) use hazards instead of failures and better information about potential causal scenarios to improve severity and likelihood estimates, and (2) change the basic definition of risk and thus its assessment. The second topic is covered in Part 2 of this white paper.

The first alternative requires the fewest changes to what is done today, i.e., use the basic risk matrix as it exists but change the way that entries are derived.

Use Hazards Rather than Failures

Some of the inaccuracy in risk matrix severity evaluations stem 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, not component or even system unreliability. Component or system reliability is not equivalent to system safety, although there are overlaps. In fact, 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 the stakeholders and then the hazards are associated with the prioritized losses. This process is easier and more straightforward than starting by attempting to prioritize 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. And, 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 (Safety Assessment Report) [Sikorsky 2012] on a Black Hawk upgrade included a failure of the aircraft's 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 by using STPA that could prevent the blade de-ice function when the APU has *not* failed. Consider the following unsafe control action:

> *UCA: The flight crew does not switch the APU (Auxiliary Power Unit) generator power ON when either GEN1 or GEN2[3] 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 [Abrecht et. al. 2016]. These are not included in the official Blackhawk SAR, but need to be factored into any risk assessment and used to develop design, testing, and operational requirements. The new scenarios for the UCA above 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 very 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) and 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 with 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

[3] Redundant APU generators

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 detailed process to identify the scenarios and not allow stopping before full consideration of these scenarios in the risk assessment. One cannot ensure completeness, of course, in any non-mathematical process, but following a rigorous process will result in reducing shortcuts and biases and in more full consideration of potential causal scenarios.

One problem in assessing likelihood is that little real information is available about the future, especially 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 drunk crawling around on a sidewalk underneath a streetlight looking for his lost wallet. The man offers to help and asks where the drunk lost the wallet. The drunk points to the other side of the street. When the man asks why he is looking in a place different than where he dropped the wallet, the drunk 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.

The scenarios generated by STPA can potentially provide better information on 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 again. 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 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 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. Factor (a) above could be controlled through redundancy and fault tolerant design. Factor (b) by interface design (as evaluated by a human factors expert). Factor (c) will also 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. We still need a way to link these factors to likelihood. A few are suggested below.

The example shown so far focuses on the interaction of the flight crew and the aircraft controls. The design of the software and hardware, of course, also must be included in the risk assessment. Current approaches to handling software, such as assigning levels of rigor to the software development, have no technical or scientific basis as mentioned above. Simply assuming that software-related risk is adequately reduced or eliminated by rigorous development is not realistic and does not reflect any research results nor 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 humans.

Consider the following example of a 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 (engine controller) 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 ICUs (inceptor control units) to properly place the collective to prevent low rotor RPM conditions.

While these STPA-generated scenarios would usually 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 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. Simply assigning a likelihood to “FCC failure” or even “Hazardous FCC behavior” is not possible. Using

the strength of potential controls would help. At the simplest level, this assessment might involve differentiating between controls that eliminate the hazard or only trying to detect and mitigate it. Note that STPA can be performed early in development and the information used to have an impact on the development process. For example, identified safety constraints could be subjected to more extensive assurance activities.

Translating Strength of Controls into Likelihood

The problem remains of associating likelihood with strength of potential controls. 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, the likelihood evaluations can be improved. At the end, the risk associated with the system during operations will be possible to evaluate with much better accuracy than currently possible.

Various strategies might be used to rank the strength of potential controls. One possible ranking (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 is perhaps too simple. A little 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 the causal factor might be able to be mitigated. This procedure may improve the results over 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."

In addition, combinations of these four types of control (listed above) might be used in likelihood estimates, e.g., design features are included to reduce or control the factor as well as operator training and procedures as a backup in case the hazard still occurs. 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 usual one 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 probabilities for these factors up out of thin air.

Specialized risk assessment processes can be developed that are appropriate for specific types of systems. Chapter 10 of Engineering a Safer World (pages 321 to 327) describes two such special approaches we have devised for past projects. The first was for a NASA contract to create and analyze architectural tradeoffs 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.

The process devised to assess risk for this architectural trade study was the following. Hazards were identified that were specific to each mission phase (e.g. launch or landing) 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 were evaluated for their impact in 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 tradeoff 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, however, severity was easier to handle than likelihood. In this case, the architectures and missions would involve things that had never been attempted before and historical data was not relevant. Instead, mitigation potential was substituted for likelihood as in the example above but in a more sophisticated way. Mitigatibility was evaluated by domain experts under the guidance of safety experts. Both the cost/difficulty of the potential mitigation strategy (in qualitative terms of low, medium, 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 used to combine mitigation factors and severity factors to come up with a final Overall Residual Safety-Risk Metric. This metric was then used in the evaluation and ranking of the potential manned space exploration architectures. A detailed example can be found in *Engineering a Safety World*.

A second example is a scheme we came up with for evaluating risk in a human-intensive NASA project involving Air Traffic Control (ATC) enhancements. This case was almost the exact opposite of that for the manned space mission design in that the system engineering problem is not to create a new or safer system but to maintain the already high level of safety built into the current system. The goal is essentially not to degrade the safety of the current system when changes are made to it. The risk analysis is then 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 very high level.

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

- *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 or no change in the controller skills, a minor change, or a significant change.
- *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 and 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 and 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 to provide better values for the criteria. The point is that thought put into what risk means in your particular project can help to 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 such as manufacturing controls, designed maintainability and the occurrence of maintenance errors, training programs, changes over time in the environment in which the system is used, 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 assumptions about the operational environment by the system designers. How realistic and accurate those assumptions are, how well those assumptions are communicated to the 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 life cycle. Castilho [Castilho 2019] has devised what he calls Active STPA, which can be used during operations to identify leading indicators of changes that increase risk.

While the use of rigorously developed causal scenarios using STPA does not avoid all the problems with standard risk matrices, it does at least provide a more rational basis for the 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.

The improvements in risk assessment described so far have involved keeping the same standard definition of risk, making changes but essentially keeping the same form for the risk matrix, and using STPA to assist in the evaluation of severity and likelihood. An alternative is to make significant changes to the definition of risk itself and therefore to its evaluation with the goal of greatly improving the results. Exploring this topic is the subject of Part 2 of this white paper.

References

Abrecht, B., Arterburn, D., Horney, D., Schneider, J. , Abel, B., and Leveson, N. (2016), A New Approach to Hazard Analysis for Rotorcraft, AHS Technical Specialists' Meeting on the Development, Affordability, and Qualification of Complex Systems, Huntsville AL, February 9–10.

Abrecht, Blake (2016), *Systems Theoretic Process Analysis Applied to an Off-Shore Supply Vessel Dynamic Positioning System*, S.M. Thesis, Aeronautics and Astronautics Dept., MIT

Castilho, Diogo Silva (2019), A Systems-based Model and Processes for Integrated Safety Management Systems (I-SMS), Ph.D. Dissertation (in process), MIT Dept. of Aeronautics and Astronautics, expected August.

Cox, Anthony (2008), What's Wrong with Risk Matrices, *Risk Analysis* 28(2):497–512.

Kahneman D, Tversky A. (1973) On the Psychology of Prediction. *Psychological Review* 80 (4):237–51.

Kahneman, D., Slovic P., and Tversky A. (1982) *Judgment under Uncertainty: Heuristics and Biases*, New York: Cambridge University Press.

Lauridsen, K., Kozine, I., Markert, F., Amendola, A., Christou, M. and Fiori, M., (2002), Assessment of *Uncertainties in Risk, 2002, Assessment of Uncertainties in Risk Analysis of Chemical Establishments,* Risø National Laboratory, Roskilde, Denmark, Risø-R-1344(EN).

Leveson, Nancy (1995), *Safeware: System Safety and Computers*, New York: Addison-Wesley

Leveson, Nancy (2012) Engineering a Safer World, Cambridge, MA: MIT Press.

Rasmussen, Jens (1997), Risk Management in a Dynamic Society: A Modeling Problem, *Safety Science* 27 (2/3):183–213.

Sikorsky Aircraft Corporation (2012), Safety Assessment Report for the UH-60M Upgrade Aircraft, Document Number SER-703655, January 3.