Impediments to Effective Safety Risk Assessment of Safety Critical Systems: An Insight into SRM Processes and Expert Aggregation

Cynthia Stephen

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Industrial and Systems Engineering

Alejandro Salado, Chair
Hanumanthrao Kannan
Brian M. Kleiner

May 21, 2020
Blacksburg, Virginia

Keywords: Safety Risk Assessment, Decision Analysis, Utility Theory, Multiattribute Utility, Expert Aggregation, Judgement Aggregation.
Impediments to Effective Safety Risk Assessment of Safety Critical Systems: An Insight into SRM Processes and Expert Aggregation

Cynthia Stephen

ABSTRACT

Safety risk assessment forms an integral part of the design and development of Safety Critical Systems. Conventionally in these systems, standards and policies have been developed to prescribe processes for safety risk assessment. These standards provide guidelines, references and structure to personnel involved in the risk assessment process. However, in some of these standards, the prescribed methods for safety decision making were found to be deficient in some respects. Two such deficiencies have been addressed in this thesis.

First, when different safety metrics are required to be combined for a safety related decision, the current practices of using safety risk matrices were found to be inconsistent with the axioms of decision theory. Second, in the safety risk assessment process, when multiple experts are consulted to provide their judgment on the severity and/or likelihood of hazards, the standards were lacking detailed guidelines for aggregating experts’ judgements. Such deficiencies could lead to misconceptions pertaining to the safety risk level of critical hazards. These misconceptions potentially give rise to inconsistent safety decisions that might ultimately result in catastrophic outcomes.

This thesis addresses both these concerns present in SRM processes. For the problem of combining safety metrics, three potential approaches have been proposed. Normative
Decision Analysis tools such as Utility Theory and Multi-attribute Utility Theory were proposed in the first and second approaches. The third approach proposes the use of a Multi-Objective Optimization technique - Pareto Analysis. For problems in Expert Aggregation, behavioral and mathematical solutions have been explored and the implications of using these methods for Safety Risk Assessment have been discussed. Two standard documents that contain the Safety Risk Management Processes of the Federal Aviation Agency (FAA) and the U.S. Navy were used to structure the case studies.

This thesis has two main contributions. First, it evaluates the use of decision analysis in safety decision process of Safety Critical Systems. It provides guidelines to decision makers on how to meaningfully use and/or combine different safety metrics in the decision process. Second, it identifies the best practices and methods of aggregating expert assessments pertaining to safety decision making.
Impediments to Effective Safety Risk Assessment of Safety Critical Systems: An Insight into SRM Processes and Expert Aggregation

Cynthia Stephen

GENERAL AUDIENCE ABSTRACT

Safety risk assessment forms an important part of the design and development of Safety Critical Systems. Safety Critical Systems are those systems whose failure could potentially result in the loss of human life. Commonly in these systems, standards and policies have been developed to prescribe processes for safety risk assessment. These standards provide guidelines, references and structure to personnel involved in the risk assessment process. However, in some of these standards, the prescribed methods for safety decision making were found to be deficient in some respects. Two such deficiencies have been addressed in this thesis.

First, when different safety metrics are required to be combined to provide information for a safety related decision, the current practices of the safety risk assessment do not yield consistent recommendations. Second, in the safety risk assessment process, often multiple experts are consulted to provide their judgment on the criticality of a potential safety risk of the system. The standards and policies that are currently being used, do not provide clear instructions on how to synthesize the judgements of multiple experts. This lack of clear guidelines could potentially lead to an incorrect final judgement on the criticality of the risk and ultimately result in choosing an improper method to reduce the safety risk.
This thesis addresses both these concerns present in safety risk assessment process of Safety Critical Systems. For the problem of combining safety metrics, three approaches have been proposed. Two of the proposed approaches make use of normative decision analysis practices and therefore the recommendations reached using these methods will be consistent with the safety objective of the decision maker. The third approach makes use of a traditional concept called -Pareto Analysis which provides a visual method to analyze the advantages and drawbacks of a given safety concern for a system.

For problems in combining the judgements of multiple experts a variety of methods was studied. The methods include group consensus and mathematical techniques and the implications of using these methods in safety risk assessment was discussed. The FAA and the U.S. Navy’s standard documents and policies were used to frame the discussions.

This thesis has two main contributions. First, it evaluates the use of Normative Decision Analysis methods in safety decision process of Safety Critical Systems. It provides guidelines to decision makers on how to meaningfully use and/or combine different safety metrics in the decision process. Second, it identifies the best practices and methods of aggregating expert assessments pertaining to safety decision making.
Dedication

To my dearest sister, Sharmila.
Acknowledgements

I would like to express my sincere gratitude to my advisor Dr. Alejandro Salado for all his support and motivation during my time at Virginia Tech. It was a privilege to learn and work under his guidance. This research would not have been possible without his encouragement and mentorship from time to time.

I would like to thank Dr. Hanumanthrao Kannan and Dr. Brian Kleiner, for sparing their time and serving on my committee. Dr. Kannan, I cannot thank you enough for your patience and all the advice and knowledge that you have shared with me over the past year. Dr. Kleiner, I sincerely appreciate your effort in reviewing my work and providing valuable suggestions.

I gratefully acknowledge Ms. Susan J. DeGuzman, Director, USN/USMC Airworthiness and CYBERSAFE Office, Naval Air Systems Command. This research was funded in by NAVAIR Contract #N00421-16-2-0001. I also acknowledge Mr. Richard E. Adams, DoN Chief Airworthiness Engineer for UAS & Targets, USN/USMC Airworthiness & Cybersafe Office, Naval Air Systems Command for his technical advice and guidance.

I take this opportunity to thank my parents Stephen and Usha, my sister Sharmila, my brother-in-law Prabhu, and my dearest niece Riya, for being my backbone throughout this journey and making me feel loved in every way possible. Thank you for your unwavering support.
Finally, I would like to thank my friends, here in Virginia Tech and back home in India. I thank Shalini, for her timely help and support over the past two years. My best friends Vanathi and Davis, who despite the distance were always there to hear me out. Thanks for believing in me and encouraging me to strive for excellence. I truly treasure our friendship.
Table of Contents

1. Relevance and Research Questions ......................................................... 1
   1.1 Introduction ...................................................................................... 1
   1.2 Safety Objectives and Safety Metrics .............................................. 3
   1.3 Measurement of Safety ................................................................. 5
   1.4 Aggregation of Expertise ............................................................. 6
   1.5 Research Questions ....................................................................... 8
   1.6 Contribution of Research ............................................................. 8
   1.7 Overview of Thesis ....................................................................... 9

2. Literature Review ................................................................................. 10
   2.1 What is Safety? ............................................................................. 10
      2.1.1 Definitions of Safety ............................................................... 12
      2.1.2 Safety Metrics in Aviation ...................................................... 14
   2.2 Safety Risk Management Process (SRMP) ...................................... 16
   2.3 Expert Aggregation ...................................................................... 22
      2.3.1 Expert Aggregation in FAA and U. S. Navy ............................. 23
      2.3.2 Discursive Dilemma and the Impossibility Result ..................... 26
      2.3.3 Potential Solutions to the Dilemma ......................................... 28
      2.3.4 Relaxing Propositional Logic ............................................... 29
      2.3.5 Endorsements for Mathematical Techniques over Behavioral
            Methods ...................................................................................... 30
      2.3.6 Mathematical Aggregation .................................................... 31
      2.3.7 Linear and Logarithmic Aggregation (Opinion Pooling) .......... 32
3. Methodology

3.1 Combining Safety Metrics

3.1.1 Illustration of Inconsistencies of Safety Risk Matrix

3.1.2 The Decision Scenario – Combining Metrics

3.1.3 Approach 1: Choose a Single Metric and Apply Utility Theory

3.1.4 Approach 2: Combine the Two Metrics using Multi-attribute Utility Theory (MAUT)

3.1.5 Approach 3: Perform a Trade Space Analysis

3.2 Expert Judgement Aggregation

3.2.1 Behavioral Aggregation

3.2.2 Mathematical Aggregation

4. Analysis and Results

4.1 Safety Risk Matrices

4.2 Approaches to Combine Safety Metrics

4.3 Approach 1: Choose a Single Metric and Apply Utility Theory

4.3.1 Using ‘Number of Fatalities’ as a Direct Value Attribute

4.3.2 Value Functions

4.3.3 Elicitation of Utility Values

4.3.4 Deriving the Single Attribute Utility Function

4.4 Approach 2: Build a Multiattribute Utility Function

4.4.1 Preference Function

4.4.2 Multi-attribute Utility Analysis
4.5 Approach 3: Perform Trade-Space Analysis.................................72
4.6 Expert Assessment Aggregation.............................................75
  4.6.1 Judgement Aggregation in Safety Risk Assessment: The Discursive
        Dilemma.................................................................74
  4.6.2 Implications of the Dilemma..........................................77
4.7 Evaluation of Potential Solutions to the Discursive Dilemma...........78
  4.7.1 Practicing CBP and PBP.............................................78
  4.7.2 Practicing Modus Tollens instead of Modus Ponens..............80
  4.7.3 Other Strategies to Evade the Dilemma..............................81
4.8 Mathematical Aggregation.....................................................84
  4.8.1 Arithmetic and Geometric Averaging for Aggregating Severity
        Rankings.................................................................84
  4.8.2 Belief Aggregation – Safety Risk Assessment......................86
  4.8.3 Belief Aggregation for Likelihood of Hazard Occurrence........88
  4.8.4 Linear Pooling.......................................................91
  4.8.5 Logarithmic Pooling..................................................93
  4.8.6 Using Likelihood as a Judgment rather than Belief...............95

5. Conclusion.................................................................................96
  5.1 Contributions, Conclusions and Recommendations....................96
  5.2 Research Limitations.........................................................100
  5.3 Future Work.........................................................................101

References....................................................................................102
Appendix A....................................................................................109
List of Figures

Figure 4.1 Small Aircraft/ Rotorcraft Safety Risk Matrix........................................47
Figure 4.2a Utility Function for a Risk Averse DM..................................................50
Figure 4.2b Utility Function for a Risk Neutral DM..................................................51
Figure 4.2c Utility Function for a Risk Seeking DM..................................................51
Figure 4.3 X- Set of Prospects ‘Number of Fatalities’.................................................57
Figure 4.4 Utility of $x_1$ Number of Fatalities.........................................................60
Figure 4.5 Utility Function for attribute X - Number of Fatalities...............................61
Figure 4.6 Preference Function P(X,Y) at $X = x_0$...................................................65
Figure 4.7 Desired Isopreference Contours.................................................................67
Figure 4.8a Preference for prospects characterized by attributes X and Y...............69
Figure 4.8b Isopreference Contours of attributes X and Y........................................70
Figure 4.8c Isopreference contours projected on X Y plane........................................72
Figure 4.9 Option Space..............................................................................................74
Figure 4.10 Probability distribution of Experts A and B..............................................88
Figure 4.11 Belief Distributions of 5 Experts...............................................................91
Figure 4.12 Linear Pooling.........................................................................................92
Figure 4.13 CDF after Linear Pooling.......................................................................93
Figure 4.14 Logarithmic Pooling...............................................................................94
Figure 4.15 CDF after Logarithmic Pooling...............................................................95
List of Tables

Table 2.1 Safety Definitions.................................................................12
Table 2.2 Common Safety Metrics used in Aviation.................................15
Table 2.3 Hazard Definitions FAA and U. S. Navy..................................18
Table 2.4 Severity Definitions of FAA and U. S. Navy............................19
Table 2.5 Illustration of the Discursive Dilemma....................................27
Table 4.1 Decoded Elements of Risk Matrix as Prospects........................48
Table 4.2 Fatality Equivalents...............................................................49
Table 4.3a Utility of Prospects for Different Risk Attitudes of DM..............52
Table 4.3b Changed Preference order of Prospects based on the Utility Functions..54
Table 4.4 Elicited Utility Values for X- Number of Fatalities......................60
Table 4.5 Preference Values for Y- Number of Accidents..........................64
Table 4.6 ‘Expected Fatalities’ and ‘Expected Accidents’ for 10 Aircraft Types....73
Table 4.7 Discursive Dilemma in Safety Risk Assessment..........................76
Table 4.8 Aggregated Judgement for CBP.............................................79
Table 4.9 Aggregated Premises for PBP................................................79
Table 4.10 Modus Tollens..................................................................81
Table 4.11 Aggregation Strategies that yields “Collectivized Reason”.........84
Table 4.12 Severity Assessments for Hazards X, Y and Z by 5 experts...........85
Table 4.13 Likelihood Definitions..........................................................87
Table 4.14 Expert Assessment of Likelihood of Hazard Causes....................89
Table 4.15 Mean Rate of failure asserted by each SME................................90
Table 5.1 Inconsistencies of Current Methods of Safety Risk Assessment ....97
Table A1 Z - Matrix Values (After Normalization)……………………………………109
Table A2.1 Fatality Probabilities for Aircraft A, B, C, ..., J…………………………109
Table A2.2 Accident Probabilities for Aircraft A, B, C, ..., J…………………………110
List of Abbreviations

CBP - Conclusion Based Procedure
DM - Decision Maker
FAA - Federal Aviation Administration
MFO - Majority Fusion Operator
PBP - Premise Based Procedure
SCS - Safety Critical System
SME - Subject Matter Expert
SMM - Safety Management Manual
SRMP - Safety Risk Management Process
Chapter 1

Relevance and Research Questions

1.1 Introduction

Normative decision analysis processes have traditionally been applied in economics and financial engineering to make strategic business choices (Kenton 2020). In recent years, the decision analysis framework has proved useful in a broad spectrum of other domains ranging from engineering design to medical decision making (Design 2001). The process has been recognized for its structured, systematic and visual approach, providing clarity of action to the decision maker (Abbas and Howard 2015).

For safety decision making of Safety Critical Systems (SCS), however, standardized safety policies and processes are common practice (Smith and Simpson 2010). A SCS is defined as a system whose failure could result in loss of life, cause damage to property and/or to the environment (Knight 2002). These standard procedures are aimed at guaranteeing at least a given level of safety. The safety of these SCSs are dictated by domain specific constraints such nature of system design, geographical area of implementation, organizational values and culture (Smith and Simpson 2010).

Organizations that employ SCSs often prescribe a Safety Management Manual (SMM) (or an equivalent formal document) that contains structured and thorough standard policies and procedures. However, a comparison of the safety decision process for aircraft systems of the FAA (Federal Aviation Administration) and the U.S. Navy described in their respective formal orders (FAA-Order8040.4B 2017) and (Navy-
OPNAVINST3500.39C 2010), reveals notable differences in the information gathering step (safety risk assessment, aggregation of expertise, etc.) of the safety decision process. These differences, and how they affect the safety decision process, are the focus of this thesis.

Two aspects of the safety decision process will be critically analyzed. The first is regarding the use of different safety metrics as objectives in the safety decision process. When different organizations using different metrics to measure safety are required to collaborate on a safety decision, the decision maker faces certain challenges such as, performing reasonable comparisons between the safety levels measured using the different metrics and/or how to combine them into a single meaningful safety metric. In this thesis, methods to overcome the discrepancies identified in the analysis are evaluated. Utility theory, a commonly used decision analysis approach, is presented as a potential solution.

The second aspect of this research pertains to the aggregation of expert assessments for uncertain elements in the decision process. Based on the standard documents used by the FAA and the U.S. Navy, several areas where expert assessment is required as input to the decision process have been identified in this thesis. Although these inputs are recognized as critical in the standards of the FAA and the U.S. Navy, the standards do not define a concrete aggregation methodology, which is left open for selection to the decision team. Practice shows that employed aggregation methodologies are often inadequate, leading to
incoherent expert aggregation. In this thesis, potential solutions to these problems are explored and assessed.

1.2 Safety Objectives and Safety Metrics

Having clear decision objectives is key to defining the value of a decision and eliminating some of the conflicts that may arise in the preferential precedence of decision prospects (Goodwin and Wright 2014). As a preliminary step to having clear safety objectives, the meaning of safety must be defined. Safety has been defined in numerous contexts and is often associated with a subjective degree of belief (Möller, Hansson et al. 2006). This subjectivity makes it difficult to define safety. Further, the linguistic imprecision of the term ‘safety’ allows for a broad set of interpretations of its meaning and might lead decision makers to contextualize safety differently, making it difficult to identify safety objectives and safety metrics to measure them (Möller, Hansson et al. 2006). For example, in formulating a safety decision problem, having an objective as broad as ‘maximize safety’ might prove difficult to comprehend and is prone to interpretations. Instead, safety objectives should be deconstructed into attributes that will enable analysis of the decision maker’s preferences and development of value measures.

When addressing the concept of safety measurement in a broad perspective, there have been several attempts to use safety metrics classified into leading and lagging indicators of safety performance (Swuste, Theunissen et al. 2016). Lagging indicators are safety measures of system’s accidents/incidents in the past (Lingard, Hallowell et al. 2017). Leading indicators are safety measures that forecast potential safety risk events (Lingard,
Hallowell et al. 2017). These safety metrics could be used as decision objectives in safety decision making. Even when limiting our discussion to aviation, there have been numerous safety risk assessment metrics that have been used as measures for safety. Examples of aviation safety metrics include “sum of expected economic losses due to fatalities and loss of equipment”, “probability of an adverse event (or occurrence of undesirable events) per unit of exposure”, “number of catastrophic events”, “number of casualties”, “number of accidents” (Speijker, Lee et al. 2011), “nominal likelihood of a mishap causing serious injury, loss of life or significant damage per flight hour” (Washington, Clothier et al. 2017), “mid-air collision rate per flight hour (Washington, Clothier et al. 2017), accident rate, incident rate” (Chang and Yeh 2004, Liou, Tzeng et al. 2007), and probability of system failure (Melnyk, Schrage et al. 2014).

In a situation with multiple stakeholders involved in the decision process, each having a different perspective on safety, the metric used to measure safety might be different based on the safety value sought by the respective stakeholder. Again, this could lead to inconsistencies in the measurement of safety, as the metrics used by the individual stakeholder groups might not enable reasonable comparisons. Therefore, it becomes necessary to develop methodologies to consistently include different safety metrics into the decision process to meaningfully measure safety. A detailed review of the meaning of safety and the corresponding metrics derived from the stated safety definition is presented in Chapter 2.
1.3 Measurement of Safety

A common connotation of safety is that a system is safe if its operation does not cause harm to humans, property or the environment (Knight 2002). The safety risk associated with the operation of the system is often used as a measure to determine the system safety. If the estimated risk is under agreed acceptable levels, then the system is said to be ‘safe’. The criticality of a safety risk is commonly defined in terms of severity of the risk and the probability of its occurrence (Navy-OPNAVINST3500.39C 2010, FAA-Order8040.4B 2017). The classification of severity and likelihood definitions are commonly provided by standard Safety Risk Management documents of the concerned system.

Traditionally, risk matrices have been used to depict system safety risk levels (Pedersen 1999). For example, if the anticipated severity of the potential safety risk is “Physical distress or injuries to persons and/or Substantial damage to aircraft/vehicle” (FAA-Order8040.6 2019) and the such risk event is “expected to occur routinely/expected to occur more than 10 times per year” (FAA-Order8040.6 2019), the criticality of the risk event would be considered High Risk when using the risk matrix of the FAA’s order 8040.6. However, using risk matrices and arbitrary scoring to assess safety risk causes discordance in the safety estimates (Rumsfeld 2002). This thesis will present alternate methods to assess safety risk that are more meaningful in the representation of safety as perceived by the safety assessor.
1.4 Aggregation of Expertise

The second objective of this research is to study inconsistencies resulting from aggregating expertise in safety decisions. In the safety decision making process in SCS, information is exchanged and gathered from several Subject Matter Experts (SMEs), described according to the respective SRMP of the system. The SMEs assess uncertain attributes in the decision problem (Vismari and Junior 2008). SMEs possess a certain level of expertise regarding the subject under investigation. The process of including SME assessment into the decision process can be viewed as a three-step approach

- **Step 1. Expert calibration**, which is the process of choosing and assigning importance/weight to the expert (Jongsawat and Premchaiswadi 2010).

- **Step 2. Expert elicitation**, which encompasses procedures to elicit the knowledge from the expert (Goodwin and Wright 2014).

- **Step 3. Expert aggregation**, which is combining the assessed information into a decision element that can be used in the decision model (van Steen 1992).

The SRMP processes of the FAA and the U.S. Navy indicate that Multiple Subject Matter Experts (SMEs) are involved in the decision process for safety risk assessment and acceptance (Navy-OPNAVINST3500.39C 2010, FAA-Order8040.4B 2017). The safety risk matrix used in the SRMPs of the FAA and the U.S. Navy depend on inputs from the SMEs. The inputs such as impact probability, severity and probability of accident occurrence are often obtained based on SMEs judgement (van Steen 1992). The SRMP provides a description regarding the responsibilities and the parties who are necessary to
be consulted in the safety decision process might be defined. However, there is no prescribed process that dictates how to aggregate the team member SME inputs in the decision-making process. This indicates that in each safety risk assessment process, a different aggregation technique might be followed without proper rationale behind the usage. To further strengthen this argument, a report published by the United States General Accounting Office states that (Rumsfeld 2002):

“Both, the FAA and the military services have internal informal networks in place among aviation safety personnel to share information. These exchanges are typically self-initiated, occur on an ad hoc basis, and are based largely on personal relationships. It is a primary means used among the military services’ safety centers to keep apprised of current aviation safety issues.”

The implications of following different aggregation schemes could result in choosing inappropriate mitigation/control strategies which are only seemingly compliant with the regulatory guidelines imposed by the organization.

The aggregation process is also sensitive to the type of data (e.g., probability distributions, qualitative statements, arbitrary scores, etc.) being aggregated. Further, the quality of aggregating expertise depends on factors such as competence of the expert, expert team composition, method of expert elicitation and behavioral/mathematical process of aggregation. Based on the type of information required in the safety decision model (opinion, belief and judgement) an aggregation methodology must be established
to provide structure to the safety decision process and ensure that the premise and judgment of the SMEs remain consistent.

1.5 Research Questions

There are several gaps in the safety decision process of SCSs that need further research and evaluation. This thesis will focus on two specific activities of the safety decision process of SCSs. The first one pertains to the meaningful inclusion of safety metrics in the decision model. The second one pertains to the aggregation of expertise. The following two research questions are addressed:

**RQ1.** How should different safety metrics in the safety decision process be aggregated to maintain consistency with the objective of ‘safety’ as perceived by the different stakeholder groups?

**RQ2.** Which expert aggregation methods preserve premise and judgement consistency in the safety assessment for safety decision making?

1.6 Contribution of Research

This research has three main contributions. First, it illustrates the inconsistencies of the current practices of safety risk assessment from a decision analysis standpoint. Mainly the use of safety risk matrix as a decision aid was scrutinized. Second, it evaluates the use of decision analysis in safety decision process of SCSs and provides guidelines to decision makers on how to meaningfully use and/or combine different safety metrics in
the decision process. Third, it identifies the best practices and methods of aggregating expert assessments pertaining to safety decision making.

1.7 Overview of Thesis

The rest of the thesis is organized as follows. Chapter 2 reviews the literature, providing evidence that safety has been interpreted and measured differently for different safety critical systems. The various definitions of safety and the frequently used safety metrics in aviation are presented. A review of the safety risk management processes followed by the FAA and the U.S. Navy, as well as their similarities and differences are discussed. Chapter 2 also presents a review of the current behavioral and mathematical expert aggregation techniques. The problems of these techniques with respect to premise and judgement inconsistencies and potential solutions to overcome those problems are elaborated. In chapter 3, the methodology, tools and data used to evaluate proposed approaches to combine/use different safety metrics in the safety decision process and the methods used to apply expert aggregation technique in safety risk assessment are described. Chapter 4 presents the implementation of the approaches to combine safety/use different safety metrics and how they differ from current methods of safety assessment. Chapter 4 also contains the evaluation of expert aggregation techniques that are best suited in safety decision scenarios. Chapter 5 presents the conclusion drawn from the results, some limitations of the research and recommendations for future work.
Chapter 2

Literature Review

This chapter provides an insight into the various definitions of safety used in SCSs. The focus is primarily concerning the view of safety in the aviation industry. A list of safety metrics that is commonly used to measure safety in aviation is provided. Next, this chapter presents a review of the SRMP followed by the FAA and the U.S. Navy. Two official orders, that serves to structure this thesis, the FAA Order 8040.4B (FAA-Order8040.4B 2017) and the Navy-OPNAVINST 3500.39C (Navy-OPNAVINST3500.39C 2010) were studied and a succinct review of their similarities and differences is provided. The second part of this chapter pertains to Expert Aggregation. An elaboration on the current practices in judgement aggregation is presented. Two broad classifications of aggregation techniques (behavioral and mathematical) are discussed. The problems associated with these techniques and solutions to overcome the problems are also explored.

2.1 What is Safety?

Safety has been defined in many different contexts and for many different domains. Informally, when attempted to define safety, a set of thought-provoking questions were framed. Questions such as, *Is safety the lack of accidents? Is it the lack of fatalities and/or serious injuries? Is the system safe if there is an accident, yet no humans are harmed? Is it safe if the probability of system failure is low?* were used to evaluate the
understanding of the concept. Answering such questions was very difficult. This paved way to investigate more into the meaning of safety.

The overall scope for safety can be divided into human, environmental and equipment safety irrespective of the context (Dezfuli, Benjamin et al. 2011). With respect to SCSs, safety has commonly been approached from a two-dimensional perspective of harm and probability of harm. Moller and Hanson however, feel that the concept of safety is under-theorised and assert that safety should be discussed in four dimensions of harm, probability, epistemic uncertainty and control (Möller, Hansson et al. 2006). Although many researchers agree that epistemic uncertainty should be included in the scope of safety definition, there has been considerable opposition for using ‘control’ to define safety (Möller, Hansson et al. 2006).

In technical contexts, safety is often viewed as the inverse of risk. It has been argued that such a definition is insufficient. This is because, the term risk in itself has various definitions and prone to subjective interpretations (Lowrance 1976). Some of the definitions for risk are, unwanted event that may or may not occur (Clothier and Walker 2015), cause of an unwanted event, probability of an unwanted event, or expected value of unwanted events (Netjasov and Janic 2008). Literature shows that this subjectivity present in understanding risk extends itself when defining safety (Lowrance 1976, Möller, Hansson et al. 2006).
To proceed with the discussion, the scope of safety will be limited to Safety Critical Systems. The much broader domain of safety in healthcare, occupational health and psychological safety is not addressed. In the next section, the various definitions of safety used in Safety Critical Systems is presented.

2.1.1 Definitions of Safety

In SCSs, measuring safety is of prime importance (Knight 2002). The preliminary step for measuring safety is to define and conceptualize its meaning. Upon surveying literature, a broad range of expressions for safety was found. Table 2.1 presents a list of safety definitions that was collected from safety critical industries, mainly aviation and aerospace. It can be seen that the definitions are moderately different.

Table 2.1 Safety Definitions

<table>
<thead>
<tr>
<th>Definition</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>“The state in which risks associated with aviation activities, related to, or in direct support of the operation of aircraft, are reduced and controlled to an acceptable level”.</td>
<td>(ICAO 2020)</td>
</tr>
<tr>
<td>“Safety is the system property or quality that is necessary and sufficient to ensure that the number of events that could be harmful to workers, the public, or the environment is acceptably low.”</td>
<td>(Hollnagel 2014)</td>
</tr>
<tr>
<td>“Safety is the quality of being unlikely to cause or occasion hurt or injury”</td>
<td>(Stevenson 2010)</td>
</tr>
<tr>
<td>“Safety is freedom from unacceptable risk”</td>
<td>(Roelen and Klompstra 2012, Li and Guldenmund 2018)</td>
</tr>
</tbody>
</table>
“The state in which the possibility of harm to persons or of property damage is reduced to, and maintained at or below, an acceptable level through a continuing process of hazard identification and safety risk management.” (Organization 2009)

“Safety is the conservation of human life and its effectiveness and the prevention of damage to items, consistent with mission requirements” (Möller, Hansson et al. 2006)

“Safety is the degree to which accidental harm is prevented, reduced and properly reacted to” (Firesmith 2003)

“Safety is an ongoing intersubjective construct not readily measured in terms of safety cultures, structures, functions, or other commonly used descriptors of technical or organizational attributes.” (Rochlin 1999)

“Safety is a state that is achieved by reducing risk to a tolerable level. Where risk is defined as the combination of the probability of occurrence of a hazardous event and the severity of the consequence” (Misumi and Sato 1999)

“Safety means no harm” (Miller 1988)

“Safety is the absence of accidents” (Tench 1985)

“A thing is safe if its attendant risks are judged to be acceptable” (Lowrance 1976)

The difficulty of defining safety has been acknowledged before. Many researchers refer to safety as an ‘abstract’ concept, which is best defined given situational constraints (Li and Guldenmund 2018). Hansson (2012) has provided guidelines on how to clearly construct the definition for safety. He proposed a ‘strategy of terminological ramification’ and three main principles that should be adopted while defining safety.

- First, a linguistic reform for the term ‘safety’ should be performed. The linguistic reform will bring about precision, conceptual need, and minimal change of safety definition. These requirements will reinforce each other and bring a greater clarity on the terms’ (safety) perception. Precision will provide a useful concept for safety
which will unify safety research and discussions on safety policies. Then, precision, and conceptual need will rigidify the change in the usage of the term ‘safety’.

- Second, the four strategies to attain conceptual clarity in ordinal language should be implemented for ‘safety’. The four strategies are description, precisification, refinement and ramification. Description consists of clarifying the established usage. Precisification consists on eliminating vagueness. Refinement consists of improving precision and ramification consists of introducing modifiers to clarify meaning of the term in different contexts.

- Finally, the definition of safety should have the three constitutive parts. The definiendum (term to be defined – in this case ‘safety’), the definiens (that which defines) and the defining connective.

The differences in the safety measurement practices of similar organizations (organizations employing operationally relatable safety critical systems) can be attributed to the differences that stems in defining safety. Therefore, to facilitate meaningful discussions on safety between two organizations, it is necessary to first clarify the meaning of safety.

2.1.2 Safety Metrics in Aviation

Having this many definitions for safety has led to the development of many metrics for measuring safety. Some of the metrics are combined with other metrics to yield stipulated ‘safety level’. Most commonly in aviation, the measure of safety is being used
interchangeably with safety risk and is calculated based on the severity and likelihood of a potential hazard. Amalberti (2001) asserts that providing a meaningful measure of safety that is based on catastrophic outcomes is difficult as the likelihood of a catastrophic event is usually below one accident per million events. Table 2.2 presents a list of commonly used safety metrics in aviation.

Table 2.2 Common Safety Metrics used in Aviation

<table>
<thead>
<tr>
<th>Safety Metric</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accident rate</td>
<td>(Chang and Yeh 2004, Shyur 2008)</td>
</tr>
<tr>
<td>Incident rate</td>
<td>(Chang and Yeh 2004, Shyur 2008)</td>
</tr>
<tr>
<td>Estimated number of fatalities per crash</td>
<td>(Netjasov and Janic 2008)</td>
</tr>
<tr>
<td>Probability of aircraft accident</td>
<td>(Netjasov and Janic 2008)</td>
</tr>
<tr>
<td>Hazard Rate</td>
<td>(Shyur, Keng et al. 2012)</td>
</tr>
<tr>
<td>Number of ground fatalities per flight hour</td>
<td>(Dalamagkidis, Valavanis et al. 2008) (Reece Clothier 2006)</td>
</tr>
<tr>
<td>Number of fatalities due to midair collisions per annum</td>
<td>(Fulton, Westcott et al. 2009)</td>
</tr>
<tr>
<td>Nominal likelihood of mishap causing serious injury, loss of life or significant damage per flight hour</td>
<td>(ADF 2009)</td>
</tr>
<tr>
<td>Mid-air collision rate per flight hour</td>
<td>(Weibel and Hansman 2004, Reece Clothier 2006)</td>
</tr>
<tr>
<td>Component failure probability or system failure that causes exposure to hazard and its consequences.</td>
<td>(Netjasov and Janic 2008)</td>
</tr>
</tbody>
</table>
The subsequent steps after utilizing these safety metrics is to define an acceptable level of risk. Agreeing on acceptable risk levels is a challenging task as it involves an agreement to be reached with multiple stakeholders having different perspectives, preferences and tradeoff that must be performed considering the available resources and cost benefit analysis (Lowrance 1976).

### 2.2 Safety Risk Management Process (SRMP)

Commonly, in SCSs there exists a Safety Management System (SMS) (Organization 2009) which includes the organization(s) safety objectives, plans, procedures responsibilities and other safety related policies. The SMS also describes a SRMP. The SRMP describes the systematic application of management policies, procedures and practices to the activities of communicating, consulting, establishing context and assessing, evaluating, treating, monitoring and reviewing risk [Definition 3.1 (ISO 2009)]. For example, consider the Safety Risk Management Process followed by the Federal Aviation Administration (FAA) and the U.S. Navy published in the FAA Order 8040.4B, 8040.6 and in U.S. NAVY OPNAVINST 3500.39C, respectively. These organizations build and operate Safety Critical Systems (SCS) where ‘safety’ is a top priority and efforts have been taken in establishing processes to maintain the safety risk of these SCSs under acceptable levels.

There are three key similarities and three differences between the aviation safety oversight systems of the FAA and the U. S. Navy (Rumsfeld 2002):
Similarities

- Internal processes for disseminating safety information
- Managing safety risks
- Certification of aircraft to meet civil standards

Differences

- Certification process that aircraft meets their *unique* safety standards
- Investigation of aircraft accidents
- Timelines and thresholds for acting on potential/identified aviation safety problems

The U.S. Navy’s SRMP (Navy-OPNAVINST3500.39C 2010), is applied to both Ship and Aircraft systems operated by the U.S. Navy. The FAA’s SRMP, FAA 8040.6 is applied in the integration of Unmanned Aircraft System (UAS) in the National Aerospace System (NAS)(FAA-Order8040.6 2019). Both these documents describe the Safety Risk Management principles and procedures that govern the systems operated by them. These documents provide definitions for the corresponding terms that are used in the safety risk assessment such as hazard, severity, likelihood, safety risk acceptance, accidents, incidents and safety related terms specific to the systems they operate. Although both these organizations operate functionally similar SCSs, there are a number of differences in their safety risk estimation procedures. In situations where the FAA and the U.S. Navy must collaborate on the safety risk assessment of a class of aircraft, these differences could potentially lead to information gaps required for the safety decision making. Some of the key differences in the safety assessment of the FAA and the U.S. Navy are presented below.
The term *Hazard* is used with slight differences by the FAA and the U.S. Navy, as shown in Table 2.3. Differences are marked with bold text. Although the definitions used are meaningful individually, certain subtle differences present might have profound effects in decision making. For instance, the U.S. Navy only considers the effects of injury, illness and death on personnel, whereas the FAA does not limit it only to personnel. Next, impact to mission accomplishment is considered a hazard by the U.S. Navy and not the FAA. Also, for the FAA, the “contributor” state is enough to be classified as a hazard. However, the U.S. Navy will classify the state as hazardous only with respect to the ultimate cause.

Table 2.3 Hazard Definition FAA and U.S. Navy (Source:(Navy-OPNAVINST3500.39C 2010, FAA-Order8040.4B 2017))

<table>
<thead>
<tr>
<th>FAA</th>
<th>U.S. NAVY</th>
</tr>
</thead>
<tbody>
<tr>
<td>“A condition that could foreseeably cause or contribute to an aircraft accident.”</td>
<td>“Any real or potential condition that can cause injury, illness, or death to personnel; damage to or loss of equipment or property; degradation of mission capability or impact to mission accomplishment; or damage to the environment. (Synonymous with the term threat.)”</td>
</tr>
<tr>
<td>Note: <em>Accident</em> is defined as an unplanned event or series of events that results in death, injury, or damage to, or loss of, equipment or property”</td>
<td></td>
</tr>
</tbody>
</table>

The safety risk assessment method also differs between the FAA and U.S. Navy, as seen in the FAA Order 8040.6, which uses 5 levels to characterize severity, and the OPNAVINST3500.39C, which uses 4 levels. Differences are shown in Table 2.4 with bold text (mainly pertaining to human subjects). The differences are not only in the
ranges, but also in the assessment framework. For example, the FAA classifies “Multiple Fatalities” as a category 1 severity. The U.S. Navy includes just 1 fatality in its category 1 severity. Multiple injuries and small number of fatalities is grouped as category 2 severity by the FAA. For the U.S. Navy, category 2 severity includes partial permanent disability, severe injury or illness. Thus, the comparison of safety assessments would not yield equivalent results.

Table 2.4 Severity Definitions of FAA and U.S. NAVY (Source:(Navy-OPNAVINST3500.39C 2010, FAA-Order8040.6 2019))

<table>
<thead>
<tr>
<th>FAA</th>
<th>U.S. NAVY</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category</strong></td>
<td><strong>Description</strong></td>
</tr>
<tr>
<td>1. Catastrophic</td>
<td><strong>Multiple Fatalities</strong> (Or fatalities to all on board) usually with the loss of aircraft/ vehicle.</td>
</tr>
<tr>
<td>2. Hazardous</td>
<td><strong>Multiple serious injuries; fatal injury to a relatively small number of persons (one or two); or a hull loss without fatalities.</strong></td>
</tr>
<tr>
<td>Level</td>
<td>Category</td>
</tr>
<tr>
<td>-------</td>
<td>----------</td>
</tr>
<tr>
<td>3. Major</td>
<td><strong>Physical distress or injuries to persons.</strong></td>
</tr>
<tr>
<td>III.</td>
<td>Degraded mission capability or unit readiness.</td>
</tr>
<tr>
<td>4. Minor</td>
<td><strong>Physical discomfort</strong> to persons.</td>
</tr>
<tr>
<td>IV.</td>
<td>Little or no adverse impact on mission capability or unit readiness.</td>
</tr>
<tr>
<td>5. Minimal</td>
<td>Negligible safety effect.</td>
</tr>
</tbody>
</table>

Subsequently, the risk matrix used to assess risk level by the FAA is a 5 x 5 matrix with a 5-scale level severity scoring system and a 5-scale level likelihood scoring system that is given both qualitative and quantitative definitions (Refer table C2. Appendix C(FAA-Order8040.6 2019)), mapped onto a final 3-level classification of safety risk, which include HIGH, MEDIUM and LOW levels. The U.S. Navy employs a 4 x 4 matrix with a 4-scale level severity scoring system and a 4-scale level likelihood scoring system (Table 2 (Navy-OPNAVINST3500.39C 2010)), mapped onto a final 5-level classification of
safety risk, which include *Critical, Serious, Moderate, Minor* and *Negligible* levels. This implies that the difference in the risk severity classification, the U.S. Navy does not monitor those risks that have negligible safety effect and those risks are automatically accepted. Upon further analysis, many of the descriptors used in the definitions of the severity and likelihood terms are subjective and are prone to personal subjective interpretations of the authoritative members (SMEs) involved in the safety risk assessment.

The problem of using rank matrices for safety risk assessment has been identified before (Pedersen 1999), yet, even in standards that have been recently published, such as the FAA order 8040.6, we see the appearance of safety risk matrices as a tool for safety assessments which are ultimately used to base safety decisions. In addition, the elements of the safety risk matrices are obtained by aggregating the assessments of SMEs. The problems present in expert aggregation techniques are elaborated in section 2.3.

The arbitrary scales and scores are used for the severity and likelihood classifications might not be the intended representations of the SME who assesses the severity and likelihood values. For instance, when considering using the severity classifications of FAA order 8040.6 for a safety decision, note how the severity ranges between category 1 to category 5 which might appeal as a linear scale. This is then open to interpretation that the value difference between a category 3 level to category 1 level event is the same as the difference between a category 5 level and a category 3 level event. However, in real life, the value for the severity classifications will decrease exponentially from category 5
to category 1. To provide a more visualizable example, the drop in preference value from no discomfort (severity 5) to physical distress (severity 3) at all will have the same drop in preference value from experiencing physical distress (Severity 3) to multiple fatalities (Severity 1). The use of arbitrary scales might mislead the reasoning/ rationale of the SME performing the safety assessment. Illustrations on how to overcome this inconsistency is discussed in the later chapters of this thesis.

Given the use of different standards, there persists a problem on how to enable reasonable safety comparisons between systems operated by these two agents (FAA and the U.S. Navy). In the situation where these two organizations must collaborate or provide assessment to a safety decision, it is highly plausible to result in miscommunication due to variations in internal safety assessment policies. One safety case report published in 2002 by the United States General Accounting Office presents how the FAA and the DoD responded to similar safety concerns (Rumsfeld 2002). The document presents the response of these organizations for four different scenarios and presents the similarities and differences in FAA’s and the Military Service under DoD’s Aviation safety oversight processes which further supports our argument that following different standards will lead to inconsistent decision processes.

2.3 Expert Aggregation

In the past, several methods have been proposed to aggregate expertise, which can be broadly classified as behavioral and mathematical (Mak, Bui et al. 1996). Behavioral methods involve information exchange and negotiation of opinions among the SMEs to
arrive at a consensus. Mathematical aggregation involves obtaining quantitative values (for the information sought) from the SMEs and applying some statistical averaging technique to combine the values (Chytka 2003). Aggregation methods have been applied in several domains and contexts (Leung and Verga 2007) and their characteristics such as robustness, traceability, prediction capabilities and accuracy have been studied (Mak, Bui et al. 1996). However, the literature is scarce in providing concrete guidelines in expert aggregation specific to safety decision making.

2.3.1 Expert Aggregation in FAA and U.S. Navy

The SRMPs of the FAA and the U.S. Navy indicate a need to incorporate the input of Subject Matter Experts (SMEs) in safety decision making. The FAA Order 8040.4B states that the safety risk analysis is carried out either by an individual or a team within a single organization. Cross-organizational teams should be formed as required by the complexity and demands of the issue at hand. The FAA and the U.S. Navy have also stated that the team of experts should have diverse skills, technical knowledge of the system and/or operation being analyzed and must also possess knowledge on safety requirements. From the documents FAA 8040.4B and the OPNAVINST 3500.39C we can infer that the SRM team members are not the decision makers. They are only responsible for providing the results of the safety assessments within their organization.

Safety Risk Assessment is an integral part of the SRMP(Navy-OPNAVINST3500.39C 2010, FAA-Order8040.4B 2017). The assessment is defined as “process of measuring or judging the value or level of something” (FAA-Order8040.4B 2017). The order further acknowledges that along different lines of businesses within the FAA there could be
different safety risk assessment processes to determine safety risk levels such as the evaluation of the probability of a fatal outcome. To overcome these differences, the FAA recommends the use of a Safety Risk Matrix.

The Safety Risk Matrix is a graphical representation of safety risk levels. The elements of the matrix are a combination of the severity of hazards and their likelihood of occurrence. The Safety risk matrix is catered according to the safety risk definitions of the FAA and the U.S. Navy’s SRM processes, which are present in the orders (FAA-Order8040.4B 2017) and the (Navy-OPNAVINST3500.39C 2010) respectively. For both these organizations, the inputs to the Safety Risk Matrix are reliant on the judgements of the SMEs. Therefore, there must be some form of judgement aggregation that takes place for the safety risk assessment in the SRMP.

Judgement aggregation is a problem for any group that tries to conclude a judgement based on sets of rationally interconnected propositions (List and Pettit 2002). However, since these problems are often associated with issues in economics, social choice and political theories, it is possible that in safety risk assessment the problems of judgement aggregation might not have even been identified in the first place.

As stated in Chapter 1, there is no concrete method that has been described on how the experts’ information is fused and applied as an element in the safety risk assessment. When a concrete procedure is not defined, it is possible that the Safety Analysts’ follow different methods to aggregate the judgements of the SMEs. Certain behavioral
aggregation methods are likely to be adopted in aggregating the SMEs assessments due to their simplicity and low computational complexity, when compared to those of mathematical methods, such as statistical averaging and Bayesian aggregation. In fact, FAA Orders 8040.4B, 8040.6 and the U.S. Navy’s OPNAVINST 3500.39C use terms such as “mutually agreed upon”, “all stakeholders have been consulted”, “peer reviewed”, or “brainstorming” to describe the aggregation process. This insight increases the confidence on the assumption that behavioral techniques might likely be followed. In addition, following different techniques to aggregation might lead to inconsistencies that arise between rationale and judgment of the teams’ aggregated safety assessment. These problems are described in the subsequent sections.

If the safety assessment team(s) are using judgement aggregation, then the impossibility theorem, the discursive dilemma (also known as the doctrinal paradox) and other problems (Phillips and Phillips 1993) might emerge and should not be ignored. When it comes to behavioral methods, structured approaches such as the Delphi technique yield better results than unstructured techniques (Gustafson, Shukla et al. 1973, Mosleh, Bier et al. 1988). Regardless of structure thought, the focus here is to illustrate how the safety risk assessment is vulnerable to problems such as the doctrinal paradox. The discursive dilemma is mainly addressed as it stands as a significant obstacle to answer the research question on how to maintain premise and conclusion consistency in judgement aggregation. In the following paragraphs, literature that addresses the problems in judgement aggregation is reviewed. Then, solutions provided in literature to overcome the problems in judgement aggregation are outlined.
2.3.2 Discursive Dilemma and the Impossibility Result

The impossibility result (List and Pettit 2002) has important implications in domains such as political theory, social epistemology, metaphysics (Weintraub 2011), economics, logic and computer science (Hartmann, Pigozzi et al. 2010). A common understanding of the impossibility statement is that, when a group of judges hold a rational set of judgements, it is not always possible for them to aggregate their judgements into a collective one in conformity to all possible constraints. When all the views of group members are required to be aggregated, it might be logically impossible to rationalize outcomes, which implies that perfect integrity is unattainable. Perfect integrity is the state of constancy and coherence (Kornhauser and Sager 2004). Coherence is the quality of being logical and consistent and constancy is the quality of being dependable (Stevenson 2010). Therefore, being unable to achieve perfect integrity implies that logical relationships cannot always be deduced from the aggregated outcome in any aggregation scheme.

An initial step in the task of aggregating judgements is that it should be distinguished from “aggregating people’s sets of credences in respect of certain propositions” (List and Pettit 2002). That is, aggregating judgements should be distinguished from the task of expert elicitation. According to List and Pettit (2002), a judgement is “an on or off affair”, meaning, the verdict of an individual concerning a proposition is binary (yes or no, true or false) and should not be mixed with belief of the individual which can be modelled as a probability distribution.
A well-known problem in judgment aggregation is the discursive dilemma (List and Pettit 2002, List and Puppe 2009, Hartmann, Pigozzi et al. 2010). A simple illustration of the discursive dilemma is shown in table 2.5. Let \( p \), \( q \) and \( r \) be three logically interconnected propositions, such that \((p \land q) \leftrightarrow r\). In a system of three judges 1, 2 and 3 each judge expresses his/ her judgement on \( p \), \( q \) and \( r \) abiding by the judgment rule \((p \land q) \leftrightarrow r\). When we observe the aggregated majority of the individual propositions, the group’s collective judgment violates the logic \((p \land q) \leftrightarrow r\). The discursive dilemma indicates that a procedure like systematic majority voting or simple majority voting on each premise cannot guarantee a rational set of collective judgements (List and Pettit 2002).

<table>
<thead>
<tr>
<th>Judge</th>
<th>( p )</th>
<th>( q )</th>
<th>( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge 1</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>Judge 2</td>
<td>True</td>
<td>False</td>
<td>False</td>
</tr>
<tr>
<td>Judge 3</td>
<td>False</td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td>Minority</td>
<td>True</td>
<td>True</td>
<td>False</td>
</tr>
</tbody>
</table>

Table 2.5 Illustration of the Discursive Dilemma

According to Arrow (2012) and List and Pettit (2002), an aggregation method does not exist that jointly satisfies the following set of requirements:

- There must be a universal domain (U). The aggregation must accept any logical judgement set (personal profile – individual expert’s judgement set).
- The judgements must be collected anonymously (A). The collective judgement must provide the same result irrespective of the order in which the individual judgement sets are collected.
- The rule of aggregation must be systematic (S). All individual propositions are treated equally (No special weighting for any individual judgement set).
The set of judgements must be consistent and deductively closed.

If all these criteria are fulfilled, then the aggregated judgement set can be termed perfect. However, Arrows Theorem (Arrow 2012) indicates that it is impossible for any aggregation method to jointly satisfy all these constraints and yet be collectively rational (List and Pettit 2002, Weintraub 2011). Such aggregation problems are applicable to most situations that require combining binary evaluations of individual voters. Section 4.6 of Chapter 4 shows how this is the case in certain safety decision making scenarios.

### 2.3.3 Potential Solutions to the Dilemma

Literature states that any aggregation procedure that satisfies the conditions of universal domain, anonymity and systematicity comes at the cost of collective rationality, consistency, or deductive closure. In addition, there can only be a procedure that roughly approximates adherence to all three principles (List and Pettit 2002). However, there are two ways that have been proposed to avoid the paradox. The Premise Based Procedure (PBP) and the Conclusion Based Procedure (CBP) (Bovens and Rabinowicz 2006, Hartmann, Pigozzi et al. 2010).

In a PBP, the individual judges express their judgement on the set of propositions $p$ and $q$ (that is, the premises in Table 1), the majority of the premise is then aggregated, and a conclusion is drawn from the aggregated premise set. In the PBP, the view taken by the majority on the conclusion will be rejected, thereby ignoring individual responsiveness and adhere to collective rationality (List and Pettit 2002).
In CBP, the judges express their final verdict on the conclusion \( r \), which abides to the logic \((p \land q) \leftrightarrow r\). Then the majority of the conclusion is aggregated as the final judgement. Thus, in CBP significance is given to the group’s conclusion; that is, maintaining individual responsiveness but violating collective rationality (List and Pettit 2002).

In the literature, there has been considerable support for utilizing PBP over CBP. An important result of Hartmann, Pigozzi et al. (2010) is that, in a voting procedure, the reasons should carry more weight than the conclusion. Mosleh, Bier et al. (1988) have also supported the use of decomposition to utilize expert’s information. It is stated that decomposition is particularly useful when different experts have more information about different aspects of the problem. Raiffa (1968) and Armstrong (1985) support decomposition, which indicates indirect support of premise based aggregation. The decomposition method is the technique where the expert is asked to respond to a series of questions on parts of the problem rather than the composite final question. Then the analyst synthesizes the responses to construct the forecast.

### 2.3.4 Relaxing Propositional Logic

One method to abide by both precepts of individual responsiveness and collective rationality is to practice modus tollens instead of modus ponens (List and Pettit 2002). Modus tollens is the propositional logic that states that if a conditional statement (“if \( p \) then \( q \)” ) is admissible, and the consequent is not true (\( \neg q \)), then the negation of the antecedent (\( \neg p \) ) can be inferred. Modus Ponens is the propositional logic that states
that if a conditional statement (“if $p$ then $q$”) is accepted, and the antecedent ($p$) holds, then the consequent ($q$) may be inferred (Yalcin 2012).

### 2.3.5 Endorsements for Mathematical Aggregation Techniques over Behavioral Methods

When we want to consider the individual team member’s competence to arrive at the (factually) right conclusion (or) when the conclusion’s (prior) probability is known, we can rely on Bayesian analysis to aggregate judgements (Hartmann, Pigozzi et al. 2010, Rufo, Pérez et al. 2012). Bayesian analysis is also recommended when there is incomplete or even misleading information. Rufo, Pérez et al. (2012) provide a Bayesian procedure to aggregate experts’ information in group decision making. Here, the belief of each expert is elicited as a multivariate prior distribution, followed by a linear or logarithmic combination methods to represent a consensus distribution. The choice of the strategy will depend on the decision maker. It has been concluded that a general unified method is achieved when they are applied to multivariate natural exponential families.

Mathematical and statistical aggregation procedures are highly prevalent, yet most of the literature is focused on aggregating probability distributions. That is, when one must aggregate beliefs then a statistical aggregation procedure has been deemed useful. There is much evidence suggesting that aggregating opinions from many experts provides more accurate results than the estimates obtained from a single expert. Also, there has been conclusive evidence that indicates that mathematical methods for aggregation generally provide better results than the behavioral methods (Von Holstein 1972, Brown 1973,

Mosleh, Bier et al. (1988) classify ‘mathematical aggregation’ as potentially beneficial compared to group consensus methods. It presents the concepts of substantive goodness and normative goodness as measures of quality for expert elicitation. Substantive goodness refers to the knowledge of goodness relative to the problem at hand. Normative goodness refers to the expert’s ability to express that knowledge in accordance with the calculus of probabilities that is a close representation of his/her actual opinion. Both types of the expertise are comparatively important. However, choosing appropriate ‘weights’ to the experts, defining proper methods for elicitation and then combining them might be overly computationally intensive (Rufo, Pérez et al. 2012) and requires expert guidance for the process itself.

Finally, it should be noted that although the aggregation techniques are classified as behavioral and mathematical, in practice, the expert aggregation process might involve some aspects of each technique (Clemen and Winkler 1999).

2.3.6 Mathematical Aggregation
Mathematical aggregation is the integration of independent assessments into a singular judgement. The level of complexity in mathematical aggregation procedures vary from simple statistical averaging to approaches based on axiomatic information gathering.
methods that incorporate the ‘expertise’ of the expert into the aggregation (Clemen and Winkler 1999, Chytka 2003). In Safety Risk assessment processes (Navy-OPNAVINST3500.39C 2010, FAA-Order8040.4B 2017) the concern is how to aggregate the ‘severity’ and ‘likelihood’ assessments of the experts. Severity (FAA-Order8040.4B 2017) has been provided as a 5 level ranked scale with 1 being highly catastrophic and 5 being minimal in causing harm. Likelihood (FAA-Order8040.4B 2017) has been given qualitative and quantitative scales from which the range of the hazard rate occurrence can be inferred. In the subsequent sections the most common mathematical aggregation techniques are discussed.

Mathematical aggregation can be generalized into linear and logarithmic opinion pools (Genest and Zidek 1986, Chytka 2003, Rufo, Pérez et al. 2012). Contextually in belief aggregation, ‘opinion’ has been used extensively and often referred as numerical statements that represent the expert’s degrees of belief on a concerned subject (Genest and Zidek 1986). The central idea in belief aggregation is to find a consensus distribution that satisfies a set of reasonable axioms (Clemen and Winkler 1999, O'Hagan, Buck et al. 2006, Rufo, Pérez et al. 2012). In the following sections, Linear, Logarithmic and Bayesian Information pooling are explained.

2.3.7 Linear and Logarithmic Aggregation (Opinion Pooling)

The general simplistic averaging technique is classified as linear opinion pooling. Here the combined final value is the weighted linear combination of the probability distribution of the belief of each expert (Genest and Zidek 1986, Clemen and Winkler
The linear opinion pooling was introduced by Stone (1961) and can be expressed as

\[ \pi(\theta) = \sum_{i=1}^{n} w_i \pi_i(\theta). \]  

(2.1)

Where \( \pi(\theta) \) is the combined probability distribution,

\( \theta \) is the quantity of interest on which the experts’ express their belief,

\( n \) represents the number of experts’ who are involved in the assessment process,

\( \pi_i(\theta) \) is the \( i^{th} \) experts’ individual probability distribution,

\( w_i \) is the weight allocated to the \( i^{th} \) expert, with weights adding up to 1;

\[ w_1 + w_2 + w_3 + \ldots + w_n = 1 \]

And \( \pi \) represents a mass function when \( \theta \) is discrete and a density function when \( \theta \) is continuous (Clemen and Winkler 1999).

The other common form of opinion pooling is the logarithmic pooling, where the combined probability distribution can be expressed as

\[ \pi_{log}(\theta) = k \prod_{i=1}^{n} \pi_i(\theta)^{w_i}. \]  

(2.2)

Where \( k \) is the normalizing constant.
In logarithmic opinion pooling, the individual belief distributions are multiplied and renormalized (Rufo, Pérez et al. 2012). In logarithmic pooling, when the weights are equal to (1/n), then $\pi_{log}(\theta)$ is proportional to the geometric mean of the individual distributions. A generalization of the linear and logarithmic opinion pools is presented in (Cooke 1991).

The linear and logarithmic opinion pooling have been used in a wide variety of domains such as medical consultation (Winkler and Poses 1993), marketing, banking, weather forecasting (Stael von Holstein 1970, Clemen and Winkler 1999) and for candidate selection of football games (Hurley and Lior 2002). In most of these applications, the weights of the experts were determined using empirical studies (Chytka 2003). These studies indicated that there was no exceptional advantage of one method over the other (Rantilla and Budescu 1999). In addition, Arrow’s impossibility result applies to these axiomatic approaches to aggregation as well (French 1983). The linear and logarithmic pooling often lead to distinctively different distributions (Rufo, Pérez et al. 2012). One seeming advantage of the logarithmic pooling over linear pooling is that, when combining density functions, the results are always unimodal; whereas linear pooling of density functions can produce combined distributions that are multimodal in nature and may cause a bias for the analyst (Chytka 2003).

Other probability combination techniques based on other statistical methods such as frequency theory have also been studied for opinion pooling (Genest and Zidek 1986).
However, for circumstances such as catastrophic disasters, where the interpretation of frequencies by the expert is stretched to the limits of plausibility, they present greater difficulties in subjectivity qualification. Furthermore, it has been established that frequency theory fails when data is sparse, unavailable, or subjected to non-sampling errors. Such problems gave endorsements to the development of Bayesian aggregation methods (Genest and Zidek 1986).

2.3.8 Advocacy for Bayesian Methods

There is extensive research in using a Bayesian aggregation scheme to support risk analysis that requires a group of experts to provide information to a decision maker (French 1983, Lindley 1985, Genest and Zidek 1986). When a prior probability distribution is available for the parameter of interest, then a Bayesian technique can be applied to update the prior distribution with a combined opinion pool (Rufo, Pérez et al. 2012). The main advantage of these Bayesian methods is that they allow for the incorporation of the ‘expertise’ level of the expert and dependencies among the experts into the aggregation model (Clemen and Winkler 1999). With \( n \) experts, providing information regarding a parameter of interest \( \theta \) and the probability distribution of \( \theta \) is known in prior \( \pi(\theta) \), then the analyst/decision maker can make use of Bayes’ Theorem to update \( \pi(\theta) \) (Clemen and Winkler 1999), using the following relationship:

\[
\pi^* = \pi(\theta | i_1, i_2, \ldots, i_n) \propto \frac{\pi(\theta)L(i_1, i_2, \ldots, i_n | \theta)}{\pi(i_1, i_2, \ldots, i_n)} \tag{2.3}
\]

Where \( L \) is the likelihood function associated with the experts’ information.
However, Bayesian-based methods have been shown to be difficult to apply in practice (Clemen and Winkler 1999) due to its computational complexity (Chytka 2003) and the difficulty of estimating the likelihood function $L$, since it must account for the prediction accuracy and bias of the individual expert.
Chapter 3

Methodology

The research methodology is structured in two sections. Section 1, Combining Safety Metrics, presents the methodology employed to answer RQ1 “How should different safety metrics in the safety decision process be aggregated to maintain consistency with the objective of ‘safety’ as perceived by the different stakeholder groups?” Section 2, Aggregating Expert Assessments, presents the methodology employed to answer RQ2 “Which expert aggregation methods preserve premise and judgement consistency in the safety assessment for safety decision making?”

3.1 Combining Safety Metrics

Utility theory was applied to the risk matrix of the FAA (FAA-Order8040.4B 2017) to investigate the deficiency of the current safety risk assessment methods. This task was done to illustrate the inconsistencies that could arise from interpreting safety risks from risk matrix. The drawback of the current methods was evaluated based on the disagreement between the outcome of the assessment and the stated qualitative standard definitions of the order (FAA-Order8040.4B 2017). Section 3.1.1 describes the methodology used to perform this task.

Next, to answer RQ1 that pertains to the use different safety metrics, a notional decision scenario was developed (Section 3.1.2). Then, three approaches to combine/use the different safety metrics was conceptualized.
Approach 1: Choose a single metric and apply utility theory.

This approach serves as a simple solution to retain consistency between the
decision makers preferences with respect to a specific safety objective and
evaluates the applicability of utility theory in safety risk assessment. Section 3.1.3
outlines the steps performed pertaining to this approach.

Approach 2: Combine the metrics using multi-attribute utility theory (MAUT).

When it is necessary to combine two different safety metrics, and if it is feasible,
multi-attribute utility theory is proposed as a potential solution. Section 3.1.4
describes the steps pertaining to this approach.

Approach 3: Perform a trade space analysis.

When combining different metrics is not feasible and/or is highly complex, then a
trade space analysis can be performed. The methodology used for this task is
described in section 3.1.5

3.1.1 Illustration of Inconsistencies of Safety Risk Matrix

In this task, the safety risk matrix of the FAA (FAA-Order8040.4B 2017) was scrutinized
and utility theory was applied to the ‘severity’ elements of the matrix. This task was used
to show that the preference function (rationale) used to develop the risk matrix was
inconsistent with the stated qualitative standard definitions. The following steps were
performed to accomplish this task.
Step 1: The matrix was decoded according to the increasing levels of severity and likelihood and was assigned a corresponding preference level. The elements of the matrix were then treated as prospects of a notional decision.

Step 2: ‘Severity’ was treated as an attribute of the prospects. Each severity level was assigned a fatality equivalent.

Step 3: Three utility functions, representing a risk averse, risk seeking, and risk neutral decision maker were applied to the fatality equivalents to analyze the changes in the preference ordering of the prospects.

3.1.2 The Decision Scenario – Combining Metrics

A notional scenario in which a situation where the FAA and the U.S. Navy must collaborate on a mission was developed. The SMEs of both the organizations will interact with a decision maker (DM) who is an external agent. The SMEs of the two organizations are requested to participate in the safety risk assessment of a specific set of newly designed aircrafts. The SMEs’ provide their assessment based on different safety metrics. Three approaches (defined above) to integrate the different assessments are investigated.

NOTE 1: The three approaches are investigated using two safety metrics, although the results are generalizable to any number of safety metrics above 1. The first metric is ‘number of fatalities’; the second metric is ‘number of accidents.’ These metrics are derived based on the definition of ‘hazard’

**NOTE 2:** In aviation, the commonly used safety metrics were found to be accident rate, incident rate, system failure rate and fatality rate. Here, for the analysis however, the ‘time’ factor’ was neglected throughout the study and the ‘number’ and/or the ‘expected number’ of the accident/fatality are considered as the safety metrics. This was done to make the metrics as meaningful direct value attributes (Abbas 2018) that can be modelled in the decision.

**NOTE 3:** The R and MATLAB programming environments were used to perform all mathematical analyses, formulations, and visual representations.

**NOTE 4:** Throughout the case, all data used were synthetic yet reasonable as per certain behavioral assumptions that were made about the decision maker.

### 3.1.3 Approach 1: Choose a Single Metric and Apply Utility Theory

In this task, utility theory is applied to a single metric. Since the derivation process of a utility function is independent of the specific safety metric on which it is applied, the metric ‘*number of fatalities*’ was chosen for demonstration in this research, without loss of generality. The task was performed using guidelines derived from (Goodwin and Wright 2014, Abbas and Howard 2015, Abbas 2018).
Step 1: A set of prospects containing the chosen metric as the outcome was defined and a preference order for the prospects was determined. The preference order was based on assumed preferences of the DM.

Step 2: The feasibility of building a value function was tested. Theoretical arguments evaluating the possibility, validity and use of a value function are discussed.

Step 3: For the elicitation of utility values from the DM, the probability equivalence rule was used. The DM was assumed to be risk averse and always prefers fewer fatalities to more.

Step 4: The utility values were then plotted on a graph. A generic utility function following Constant Absolute Risk Aversion (CARA) was employed.

Step 5: The ‘Mosaic’ data analysis package available in the R programming environment was used to identify functional parameters to fit data points and the utility function was derived.
3.1.4 Approach 2: Combine the Two Metrics using Multi-attribute Utility Theory (MAUT)

In this task, two safety metrics were combined into a single one. Number of accidents and number of fatalities are used as the safety metrics that will be combined into a single multi-attribute utility.

NOTE: These two metrics were considered as there is a level of dependency between the two metrics. A fatality can occur only when an accident has occurred. However, an accident can occur with no recorded fatalities.

Both these metrics are now treated as attributes of the decision prospects. The following steps were performed to complete this task.

Step 1: The first step was to construct a combined preference function that captures the interaction between the two attributes. The generalized version of the Cobb-Douglas production function from economics was employed. This was chosen as it allowed to vary the degree of interaction between the two attributes. The degree of interaction is important as it reflects the preference of the decision maker with respect to each attribute.

Step 2: Next, to construct the combined utility function, properties such as mutual utility independence and attribute dominance between the two attributes was verified based on the DMs assumed preferences.
Step 3: Iso-preference contours were constructed to identify the prospects that have the same preference values. The ‘contourf’ function available in MATLAB was employed to identify the iso-preference contour space.

Step 4: The multi-attribute utility values were derived from the contour plot based on contour tracing and referencing to the one-dimensional utility analysis performed in task 3.1.3.

3.1.5 Approach 3: Perform a Trade Space Analysis

In this task, a set of 10 aircrafts that were assumed to have different safety risk profiles was considered for the case study. Their probabilities for fatalities and accidents were generated synthetically. The following steps were performed to accomplish this task.

Step 1: The ‘Expected Fatalities’ and ‘Expected Accidents’ values were calculated for each aircraft. The ‘Expected Fatalities’ values for a specific aircraft are given by the sum of the product of the number of fatalities and its corresponding probability of occurrence. Similarly, the ‘Expected Accidents’ for an aircraft is given by the sum of the product of the number of accidents and its corresponding probability of occurrence.

Step 2: The option space was generated which consisted of the ‘Expected Fatalities’ and ‘Expected Accidents’ for all 10 aircrafts. The option space was
represented using the ‘plotly’ data visualization package available in the R programming environment.

**Step 3:** Finally, an analysis of the option space was presented, evaluating the Pareto optimal and dominated solutions.

### 3.2 Expert Judgement Aggregation

As stated in Chapter 1, in the SRMP of the FAA and the U.S. Navy, expert judgement is needed for severity and likelihood assessments of the identified hazards (Navy-OPNAVINST3500.39C 2010, FAA-Order8040.4B 2017). Chapter 2 discussed the various behavioral and mathematical aggregation schemes available in literature. In Chapter 4, these behavioral and mathematical aggregation schemes are applied to notional cases involving a panel of judges casting their judgement on the severity and likelihood of hazards. From reviewing the documents (FAA-Order8040.4B 2017) and (Navy-OPNAVINST3500.39C 2010) it was not evident what aggregation methodology is being practiced in the FAA and the U.S. Navy. Therefore, both behavioral and mathematical aggregation schemes were studied for their implications in safety risk assessment. Further solutions present in literature (elaborated in Chapter 2) were applied to the case study to tackle the expert aggregation problems. A conceptual argument has been presented to verify if those solutions were applicable to safety risk assessment. This analysis helped to answer RQ2 “Which expert aggregation methods preserve premise and judgement consistency in the safety assessment for safety decision making?” The methodology used to apply these aggregation schemes is described below.
3.2.1 Behavioral Aggregation

A notional case study involving a panel of three experts was presented to demonstrate the possibility of the discursive dilemma (a problem giving rise to premise and judgement inconsistencies) in a safety decision making scenario. Their judgements on severity and likelihood of an assumed hazard were synthetically established and the safety risk was determined through a logical relationship between safety risk, severity and likelihood. The logical relationship was derived with reference to the safety risk matrix presented in (FAA-Order8040.4B 2017). The solutions to overcome the dilemma presented in Chapter 2 were applied to this case and a ‘what if’ scenario analysis analogous to the practices in project management (Snyder 2014) was performed. The critical impacts of the potential ‘solution’ methodology were discussed.

3.2.2 Mathematical Aggregation

Two mathematical aggregation techniques, namely arithmetic and geometric averaging, were applied to a notional case to aggregate the severity judgements of 5 experts. A theoretical evaluation was performed on the aggregated judgements. The concepts such as ‘divergence’ of the SME judgements, the ‘Cassandra syndrome’ and their potential impacts on the assessments were discussed. To illustrate the aggregation of ‘likelihood’ judgements, the linear and logarithmic opinion pooling mechanisms described in Chapter 2 were applied to the same case and the feasibility of the aggregations were verified. The MATLAB programming environment was used to model the aggregated the probability distributions of the experts’ judgements.
This chapter presents the analysis and evaluations that were performed to answer the research questions RQ1 and RQ2. The first part of this chapter shows the discrepancies of the current practices of safety risk assessment and how they are incompliant with certain axioms of decision theory. Following this, the implementation of three proposed approaches to meaningfully combine/use different safety metrics has been presented. The latter part of this chapter illustrates various judgement aggregation methods (elaborated in Chapter 2) implemented in safety risk assessment. Finally, the implications of the various techniques of expert aggregation are discussed.

4.1 Safety Risk Matrices

As previously discussed, risk matrices are prevalent in safety risk assessment to infer the criticality of a potential hazard. The criticality of the safety risk is calculated as a function of the severity level and likelihood of the hazard (Navy-OPNAVINST3500.39C 2010, FAA-Order8040.4B 2017). The problems of using risk matrices are plentiful (Pedersen 1999). Here, the inadequacy of the safety risk matrix to reflect the preferences of a DM or an SME is shown. The incompliance of the safety risk matrix with decision theory is also illustrated. The risk matrix of the FAA shown in Fig 4.1 is used as reference for the following case study.
From a decision theory standpoint, the elements of the risk matrix (Fig 4.1) effectively represent a set of prospects. Each element is a single prospect with severity as an attribute and the corresponding likelihood as the estimated probability of occurrence. The preference order ranking of the prospects can be inferred from the colour codes presented to classify the prospects as High risk, Medium risk and Low risk. All prospects that are categorized with the same colour are presumed to have equivalent preference. The prospects categorized as high risk have the least preference and the prospects categorized as low risk have the highest preference:

Low Risk $\succ$ Medium Risk $\succ$ High Risk
The 25 elements of the risk matrix are decomposed as prospects and are assigned a preference ranking. The elements of the risk matrix decoded as prospects of a decision is given in table 4.1.

<table>
<thead>
<tr>
<th>Risk Level</th>
<th>Prospect Preference</th>
<th>Severity</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Risk</td>
<td>5E Minimal</td>
<td>Ext. Improbable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5D Minimal</td>
<td>Ext. Remote</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4E Minor</td>
<td>Ext. Improbable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5C Minimal</td>
<td>Remote</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3E Major</td>
<td>Ext. Improbable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5B Minimal</td>
<td>Probable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4D Minor</td>
<td>Ext. Remote</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2E Hazardous</td>
<td>Ext. Improbable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5A Minimal</td>
<td>Frequent</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4C Minor</td>
<td>Remote</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3D Major</td>
<td>Ext. Remote</td>
<td></td>
</tr>
<tr>
<td>Medium Risk</td>
<td>4B Minor</td>
<td>Probable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2D Hazardous</td>
<td>Ext. Remote</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1E Catastrophic</td>
<td>Ext. Improbable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3C Major</td>
<td>Remote</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4A Minor</td>
<td>Frequent</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3B Major</td>
<td>Probable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2C Hazardous</td>
<td>Remote</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1D Catastrophic</td>
<td>Ext. Remote</td>
<td></td>
</tr>
<tr>
<td>High Risk</td>
<td>3A Major</td>
<td>Frequent</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1C Catastrophic</td>
<td>Remote</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2B Hazardous</td>
<td>Probable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2A Hazardous</td>
<td>Frequent</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1B Catastrophic</td>
<td>Probable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1A Catastrophic</td>
<td>Frequent</td>
<td></td>
</tr>
</tbody>
</table>

The ‘severity’ of the elements are analyzed to be a composite of many factors such as fatalities, accidents, property damage, equipment loss etc. that was inferred from the severity definitions given in (FAA-Order8040.4B 2017). However, the method of combining both aspects of fatalities and accidents into a single severity classification might not be consistent. A detailed explanation of this problem is given in section 4.4.
Hence, only the number of fatalities was considered in this case to calculate the utility of the prospects.

Each severity level was assigned an arbitrary yet reasonable value as fatality equivalent. A maximum number of 10 fatalities was considered for the case study. Therefore, the scenario of encountering 10 fatalities has been classified as severity level 1- Catastrophic. The assumed fatality equivalents for the severity classifications is presented in table 4.2

<table>
<thead>
<tr>
<th>Severity</th>
<th>Fatality Equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1- Catastrophic</td>
<td>10</td>
</tr>
<tr>
<td>2- Hazardous</td>
<td>3</td>
</tr>
<tr>
<td>3- Major</td>
<td>0.5</td>
</tr>
<tr>
<td>4- Minor</td>
<td>0.001</td>
</tr>
<tr>
<td>5- Minimal</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

The next step was to calculate the utility value of the prospects. The utility value can be defined as the indifference probability $p$ to which the DM will be indifferent to receiving that prospect for certain or a binary deal of receiving the best prospect with a probability $p$ and the worst prospect with a probability $1 – p$ (Goodwin and Wright 2014). Three utility functions, one for a risk averse DM, one for a risk neutral DM and one for a risk seeking DM were derived. The procedure followed for deriving the utility functions is identical to the procedure explained in section 4.3.3.
The three utility functions used to calculate the utility of the prospects are presented in equations 4.1, 4.2 and 4.3. The fatality equivalents given in Table 4.2 were substituted in these equations to obtain the specific prospects utility values. It was observed that the utilities of the prospects varied diversely according to the severity classifications. The obtained utility values for each prospect is presented in Appendix Table A1.a. The graphical representation of the three utility functions following equations 4.1, 4.2 and 4.3 is presented in Fig 4.2a, 4.2b and 4.2c respectively.

Risk Averse DM

\[ U(\text{Severity}) = 0.037 + 0.967 e^{-0.866(\text{Severity})} \] (4.1)

Risk Neutral DM

\[ U(\text{Severity}) = 0.923 - 0.097(\text{Severity}) \] (4.2)

Risk Seeking DM

\[ U(\text{Severity}) = 0.974 - 0.008 e^{0.474(\text{Severity})} \] (4.3)

Fig 4.2a Utility Function for a Risk Averse DM
Fig 4.2b Utility Function for a Risk Neutral DM

Fig 4.2c Utility Function for a Risk Seeking DM
Table 4.3a Utility of Prospects for Different Risk Attitudes of DM

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Risk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3E</td>
<td>Major</td>
<td>0.5</td>
<td>0.664</td>
<td>0.877</td>
<td>0.964</td>
<td></td>
</tr>
<tr>
<td>5B</td>
<td>Minimal</td>
<td>0.001</td>
<td>1.004</td>
<td>0.926</td>
<td>0.966</td>
<td></td>
</tr>
<tr>
<td>5D</td>
<td>Minimal</td>
<td>0.0001</td>
<td>1.004</td>
<td>0.926</td>
<td>0.966</td>
<td></td>
</tr>
<tr>
<td>4E</td>
<td>Minor</td>
<td>0.001</td>
<td>1.003</td>
<td>0.926</td>
<td>0.966</td>
<td></td>
</tr>
<tr>
<td>5C</td>
<td>Minimal</td>
<td>0.0001</td>
<td>1.004</td>
<td>0.926</td>
<td>0.966</td>
<td></td>
</tr>
<tr>
<td>2E</td>
<td>Hazardous</td>
<td>3</td>
<td>0.109</td>
<td>0.634</td>
<td>0.941</td>
<td></td>
</tr>
<tr>
<td>5A</td>
<td>Minimal</td>
<td>0.0001</td>
<td>1.004</td>
<td>0.926</td>
<td>0.966</td>
<td></td>
</tr>
<tr>
<td>4C</td>
<td>Minor</td>
<td>0.001</td>
<td>1.003</td>
<td>0.926</td>
<td>0.966</td>
<td></td>
</tr>
<tr>
<td>3D</td>
<td>Major</td>
<td>0.5</td>
<td>0.664</td>
<td>0.877</td>
<td>0.964</td>
<td></td>
</tr>
<tr>
<td>Medium Risk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4B</td>
<td>Minor</td>
<td>0.001</td>
<td>1.003</td>
<td>0.926</td>
<td>0.966</td>
<td></td>
</tr>
<tr>
<td>2D</td>
<td>Hazardous</td>
<td>3</td>
<td>0.109</td>
<td>0.634</td>
<td>0.941</td>
<td></td>
</tr>
<tr>
<td>1E</td>
<td>Catastrophic</td>
<td>10</td>
<td>0.037</td>
<td>0</td>
<td>0.059</td>
<td></td>
</tr>
<tr>
<td>3C</td>
<td>Major</td>
<td>0.5</td>
<td>0.664</td>
<td>0.877</td>
<td>0.964</td>
<td></td>
</tr>
<tr>
<td>4A</td>
<td>Minor</td>
<td>0.001</td>
<td>1.003</td>
<td>0.926</td>
<td>0.966</td>
<td></td>
</tr>
<tr>
<td>3B</td>
<td>Major</td>
<td>0.5</td>
<td>0.664</td>
<td>0.877</td>
<td>0.964</td>
<td></td>
</tr>
<tr>
<td>2C</td>
<td>Hazardous</td>
<td>3</td>
<td>0.109</td>
<td>0.634</td>
<td>0.941</td>
<td></td>
</tr>
<tr>
<td>1D</td>
<td>Catastrophic</td>
<td>10</td>
<td>0.037</td>
<td>0</td>
<td>0.059</td>
<td></td>
</tr>
<tr>
<td>High Risk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3A</td>
<td>Major</td>
<td>0.5</td>
<td>0.664</td>
<td>0.877</td>
<td>0.964</td>
<td></td>
</tr>
<tr>
<td>1C</td>
<td>Catastrophic</td>
<td>10</td>
<td>0.037</td>
<td>0</td>
<td>0.059</td>
<td></td>
</tr>
<tr>
<td>2B</td>
<td>Hazardous</td>
<td>3</td>
<td>0.109</td>
<td>0.634</td>
<td>0.941</td>
<td></td>
</tr>
<tr>
<td>2A</td>
<td>Hazardous</td>
<td>3</td>
<td>0.109</td>
<td>0.634</td>
<td>0.941</td>
<td></td>
</tr>
<tr>
<td>1B</td>
<td>Catastrophic</td>
<td>10</td>
<td>0.037</td>
<td>0</td>
<td>0.059</td>
<td></td>
</tr>
<tr>
<td>1A</td>
<td>Catastrophic</td>
<td>10</td>
<td>0.037</td>
<td>0</td>
<td>0.059</td>
<td></td>
</tr>
</tbody>
</table>

When substituting these fatality equivalents in the three utility functions, the preference order ranking of the prospect’s changed. This indicates that the rationale behind the
preference classification of the elements of the risk matrix are flawed and do not abide by the axioms of decision theory. The changed order of the prospects is given in table 4.3b.

From table 4.3b it can be seen that risks that have been classified as medium, actually pose more critical considering preference of the ‘severity’ level. Also, some risks that was classified as high previously actually have a lower preference than what is inferred from the risk matrix. Thus, the safety risk assessments using the risk matrix could potentially lead to inconsistent decisions. Also, generalizing the combination of ‘severity’ and ‘likelihood’ into ‘High’, ‘Medium’ and ‘Low’ risks implies that the same course of mitigation/control strategies applies for the specific categorized risk level. However, the actual ordering of the ‘severity’ cases varies drastically, if the ‘Expected Fatalities’ for each ‘severity’ level was considered.
The severity definitions of both the FAA and the Navy (Navy-OPNAVINST3500.39C 2010, FAA-Order8040.4B 2017), include both fatalities and accident metrics. This could be one of the reasons for the inadequacy of the risk matrix method of safety risk assessment. In the following sections, three approaches are proposed to combine different safety metrics into the safety decision process which are consistent with the axioms of decision theory.

Table 4.3b Changed Preference Order of Prospects based on the Utility Functions

<table>
<thead>
<tr>
<th>Risk Level</th>
<th>Preference Order of Prospects before Utility Theory</th>
<th>Preference Order of Prospects after applying Utility Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Risk</td>
<td>5E 5D 4E 5C 3E 5B 4C 2E 5A 4B 4C 3D</td>
<td>5E 5D 5C 5B 5A 4E 4D 4C 4B 4A 3E</td>
</tr>
</tbody>
</table>
4.2 Approaches to Combine Safety Metrics

The first research question was ‘How should different safety metrics in the safety decision process be aggregated to maintain consistency with the objective of ‘safety’ as perceived by the different stakeholder groups?’ In order to answer this question, a notional case was developed based on the review of the two documents (FAA-Order8040.4B 2017) and (Navy-OPNAVINST3500.39C 2010). From analysis of the SRMP processes of the FAA and the Navy it was established in Chapter 2, that it is highly plausible that these organizations use different metrics to assess safety risk levels of potential hazards. For this case study it is assumed that the FAA uses the metric, ‘Expected number of accidents’ and the U.S. Navy uses ‘Expected number of fatalities.’

This is because, the FAA follows the hazard definition as, “A hazard is a condition that could foreseeably cause or contribute to an aircraft accident” (FAA-Order8040.4B 2017). This gives a forefront that the FAA places its safety objective based on the number of accidents encountered by an aircraft. On the other hand the U.S. Navy follows the hazard definition as “Any real or potential condition that can cause injury, illness, or death to personnel; damage to or loss of equipment or property; degradation of mission capability or impact to mission accomplishment; or damage to the environment. (Synonymous with the term threat.)” (Navy-OPNAVINST3500.39C 2010). Through this, it can be asserted that the Navy could plausibly place its safety objective based on the number of fatalities.
When the SMEs of these organizations provide their safety assessment to the external DM (Decision Maker), he/she is now challenged on how to combine these safety objectives to meaningfully assess safety levels and how to use that assessment as inputs to base his/her decision on. In Chapter 3, the methodology description for three potential approaches to tackle this challenge was presented. Here the analysis and the implantation of the three approaches is described.

4.3 Approach 1: Choose a Single Metric and Apply Utility Theory

The first approach suggests the use of a single metric to retain consistency with the safety objective. This approach facilitates the evaluation of the risk attitude and preference order of the prospects of the DM. The factors that contribute directly to the preference order of the prospects in the decision are the direct value attributes (Abbas 2018). The choice of proper direct value attributes is significant to quality decision making. Using arbitrary metrics (in this case, safety performance metrics) as direct value attributes (“an attribute is a factor that characterizes the preferences for the prospects”) in the decision-making process has been associated with low-quality decision making (Abbas 2018). It can be heavily debated whether ‘expected number of fatalities’ and ‘expected number of accidents’ can be called as direct value attributes of safety. However, that discussion is beyond the scope of this research and the case is pursued by assuming these metrics are direct value attributes.

4.3.1 Using ‘Number of Fatalities’ as a Direct Value Attribute

The assumed safety metric of the U.S. Navy, ‘number of fatalities’, can take the range of values from 0 to 10 for a specified number of flight hours for a specified unmanned
aircraft. Let $X$ denote the set of prospects characterized by the attribute -fatalities ranging from 0 to 10 fatalities.

$$X = \{x_0, x_1, x_2, \ldots, x_{10}\}$$

$x_0$ denotes the prospect of resulting in 0 fatalities, $x_1$ denotes the prospect of resulting in 1 fatality and so on up-to $x_{10}$ denoting the prospect of resulting in 10 fatalities. The DM’s objective is to maximize safety, so, it is assumed that he/she will prefer a lesser number of fatalities to more. The DM has the highest preference to have 0 fatalities and the least preference is for the prospect which results in 10 fatalities, and hence the preference order for the prospects can be written as:

$$x_0 \succ x_1 \succ x_2 \succ \cdots \succ x_{10}$$

When attempted to build a preference function for the prospects, it was a strictly decreasing function as per the assumption that lesser fatalities are always preferred to more in the safety decision frame. A set of transformations can be applied to the preference function to build a value function, from which the utility function can be derived (Abbas 2018). The feasibility and necessity of building a value function is discussed in the subsequent sections.

### 4.3.2 Value Function

Two rules are important to define a value measure: (1) The value measure should be defined on an absolute scale, and (2) differences in the value measures between the prospects should be meaningful. Several different types of measures such as arbitrary
constructed scores, weighted scores, or scaled scores using some arbitrarily developed logic could be used. However, research shows that assigning arbitrary scores and eliciting values are highly prone to inconsistencies in reflecting the actual value desired by the decision maker. Therefore, it is common practice to use monetary amounts to determine the value measures, which is familiar, fungible, divisible, alienable, and simple (Abbas and Howard 2015). Specifically, the indifference selling price of the prospect is often suggested as the value of a prospect (Abbas 2018).

A potential problem of using a monetary measure to value fatalities stems from ethical issues. While people are comfortable using monetary measures to characterize prospects involving monetarily quantifiable outcomes, such as mission profit or estimated damage to property, monetizing (or valuing in general) human life often causes major ethical dilemmas for governments and powerful figures whose decisions could impact the lives of humans. However, research has also shown that valuing life is essential to guaranteeing meeting ethical principles when engineering systems, given that they cannot achieve 100% safety (Salado and Katz 2019). Deriving a value function is however highly complex when human life is involved, and the derived value function may not yield the true preferences of the decision maker.

4.3.3 Elicitation of Utility Values

Utility elicitation is the process of eliciting utility values for the prospects from the decision maker. There are many approaches to utility elicitation such as the probability equivalence, certainty equivalence, or preference comparison (Farquhar 1984,
Chajewska, Koller et al. 2000). The elicitation process is important because utilities are highly sensitive to the method used to elicit them (Goodwin and Wright 2014). Identifying the optimal method of utility elicitation is beyond the scope of this research. Instead, the probability equivalence approach is used here, on account of its prevalent usage and simplicity (Goodwin and Wright 2014). Using the probability equivalence approach, the utility value for a prospect could be defined as the probability \( p \) that makes the decision maker indifferent to receiving that prospect for certain or receiving a binary deal that provides a probability \( p \) for receiving the most preferred prospect and a probability of \( 1-p \) of receiving the least preferred prospect. In this way, the utility function inherently incorporates the risk attitude of the decision maker. It is assumed that the utilities derived are consistent with the preferences of the decision maker.

Using the set of prospects \( X \) defined earlier, representing ‘number of fatalities,’ a utility of 1 is assigned to the most preferred prospect and a utility of 0 to the least preferred prospect:

\[
U(x_0) = 1 \\
U(x_{10}) = 0
\]

The remaining prospects have been synthetically generated assuming a risk averse decision maker. For example, suppose that the decision maker is indifferent between having 0 fatalities or a deal where he/she will receive a 0.45 probability of resulting in 1 fatality and 0.55 probability of resulting in 10 fatalities as shown below. The utility calculation for 1 fatality is based on the prospect shown in Fig. 4.4 and is given by,

\[
U(x_1) = 0.55 \cdot U(x_0) + 0.45 \cdot U(x_{10})
\]
= 0.55(1) + 0.45(0)

= 0.55

Table 4.4 lists the utilities that are used in this study.

Fig 4.4 Utility of x₁ Number of Fatalities

Table 4.4 Elicited Utility Values for X- Number of Fatalities

<table>
<thead>
<tr>
<th>Prospect (X Number of Fatalities)</th>
<th>Utility U(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x₀</td>
<td>1</td>
</tr>
<tr>
<td>x₁</td>
<td>0.55</td>
</tr>
<tr>
<td>x₂</td>
<td>0.35</td>
</tr>
<tr>
<td>x₃</td>
<td>0.28</td>
</tr>
<tr>
<td>x₄</td>
<td>0.21</td>
</tr>
<tr>
<td>x₅</td>
<td>0.18</td>
</tr>
<tr>
<td>x₆</td>
<td>0.15</td>
</tr>
<tr>
<td>x₇</td>
<td>0.1</td>
</tr>
<tr>
<td>x₈</td>
<td>0.05</td>
</tr>
<tr>
<td>x₉</td>
<td>0.04</td>
</tr>
<tr>
<td>x₁₀</td>
<td>0</td>
</tr>
</tbody>
</table>
4.3.4 Deriving the Single Attribute Utility Function

The elicited utility values were plotted against the number of fatalities. The red points marked in Fig 4.5 show the utility values for the respective number of fatalities. In order to obtain a utility function from the elicited utilities, a general form of an exponential function following Constant Absolute Risk Aversion (CARA) was used (Abbas 2018). The general form of the utility function considered is given by:

\[ U(X) = a + be^{-\gamma x} \]  (4.4)

Where, \( a, b, \gamma \) are constants and \( b \neq 0 \).

\( \gamma \) is the risk aversion co-efficient.

By fitting this function to the elicited utility values, the parameters \( a, b \) and \( \gamma \) were obtained as 0.05697, 0.911 and 0.5066, respectively. Thus, the derived utility function for the safety metric ‘number of fatalities’ can be expressed as,

\[ U(X) = 0.05697 + 0.911e^{0.5066x} \]  (4.5)

The utility function is represented in Fig 4.5.
After asserting the probabilities for the prospects, the expected utility can be calculated. As stated before, a rational decision maker should choose the alternative with the highest expected utility. By using the utility theory approach, coherence - logic and consistency can be maintained in the decision process (Goodwin and Wright 2014).

4.4 Approach 2: Build a Multiattribute Utility Function

When two or more attributes are used to characterize safety, then a multi-attribute utility function can be used to determine the combined utility and analyze trade-off preferences between the two attributes.

Let each prospect of the decision be characterized by the attributes $X$ and $Y$, where $X$ represents the set ‘Number of Fatalities’ and $Y$ represents the set ‘Number of Accidents’:
\[ X = \{ x_0^*, x_1, x_2, ..., x_{10} \} \]
\[ Y = \{ y_0^*, y_1, y_2, ..., y_{10} \} \]

The symbol * is used to denote the most preferred level for that attribute.

A preference function can be used to capture the trade-off between the two attributes under consideration. In this example, the procedure of creating the multi-attribute preference function was leveraged from Abbas (2018). It was assumed that less of an attribute is preferred to more for both \( X \) and \( Y \). The derivation of the preference function is described in the next section.

### 4.4.1 Preference Function

The preference function satisfies the following relationship

\[ P(x_{\min}, y_{\min}) = 1 \]

and

\[ P(x_{\max}, y_{\max}) = 0 \]

To determine the Preference Function that captures the interaction between the two attributes, two cases are defined: (i) At \( X = x_0 \) and (ii) for \( X > x_0 \).

**i. At \( X = x_0 \)**

When the number of fatalities is 0, the preference for the number of accidents (attribute \( Y \)) can be elicited using a similar procedure as the single attribute utility assessment presented in section 4.3.3. Here the preference \( P(x_0, Y) \) is the utility \( U(Y) \) and can simply be called as the one-dimensional utility assessment of \( Y \). That is \( U(x_0, y) = U(Y) \). The
elicited preference values (utilities) used in this example, for which risk aversion was assumed, are given in Table 4.5

Table 4.5 Preference Values for Y Number of Accidents

<table>
<thead>
<tr>
<th>Y Number of Accidents</th>
<th>P (X, Y) at X=x₀</th>
</tr>
</thead>
<tbody>
<tr>
<td>y₀</td>
<td>1</td>
</tr>
<tr>
<td>y₁</td>
<td>0.75</td>
</tr>
<tr>
<td>y₂</td>
<td>0.68</td>
</tr>
<tr>
<td>y₃</td>
<td>0.45</td>
</tr>
<tr>
<td>y₄</td>
<td>0.31</td>
</tr>
<tr>
<td>y₅</td>
<td>0.28</td>
</tr>
<tr>
<td>y₆</td>
<td>0.15</td>
</tr>
<tr>
<td>y₇</td>
<td>0.1</td>
</tr>
<tr>
<td>y₈</td>
<td>0.05</td>
</tr>
<tr>
<td>y₉</td>
<td>0.02</td>
</tr>
<tr>
<td>y₁₀</td>
<td>0</td>
</tr>
</tbody>
</table>

The preference function at $X = x₀$, for which risk aversion was assumed, is given by:

$$P(x₀, y) = K + Ce^{-\phi y}$$  \hspace{1cm} (4.7)

Where $K$, $C$ and $-\phi$ are constant. With $K = -0.1642$, $C = 1.166$, and $\phi = -0.2$ the preference function takes the form shown in Fig 4.6.

64
Note: Since the most preferred scenario is given by 0 fatalities and 0 accidents, the preference value will be the maximum (1) for $X = x_0$ and $Y = y_0$, hence:

$$P(x_0, y_0) = 1$$ (4.8)

ii. For $X > x_0$

When the number of fatalities is 1 or more, the different combinations of the prospects that have the same preference value were grouped together (it should be noted that this is possible only when there is a minimum of 1 accident):

$$P(X_i, Y_i), \forall i \in \{1, 2, \ldots, 10\}$$
The relation, $P(x,y) = a$, defines a set of points known as the level set, where $a$ represents the level set. The level set, $a$, of a preference function is determined by the set of prospects that satisfy $P(x,y) = a$, i.e.,

$$\{(x,y) : P(x,y) = a\} \quad (4.9)$$

Prospects on the same level set have the same preference and are said to lie on the same iso-preference contour. In order to determine the shape of the iso-preference contour, and since the preference is constant across a contour, the total change in the preference must be zero. Therefore,

$$dP(x,y) = 0 \quad (4.10)$$

This implies that,

$$\frac{\partial}{\partial x} P(x,y) dx + \frac{\partial}{\partial y} P(x,y) dy = 0 \quad (4.11)$$

Since less of an attribute (Number of fatalities & Number of accidents) is preferred to more

$$\frac{\partial}{\partial x} P(x,y) < 0 \quad \text{and} \quad \frac{\partial}{\partial y} P(x,y) < 0$$

Rearranging equation 4.11, we get

$$\frac{dy}{dx} = -\frac{\partial P(x,y)}{\partial x} < 0 \quad (4.12)$$

Therefore, the slopes of the iso-preference contours must be negative. The desired preference contours appear as Figure 4.7. The contour plot was obtained using a mirroring function which inverts the contour lines about the origin. The mirroring was
done to represent the preferential ordering of the two attributes $X$ and $Y$, since with respect to both attributes, it was assumed that less is preferred to more.

![Desired Iso-preference Contours](image)

Fig 4.7 Desired Iso-preference Contours

In order to assess the trade-offs among the attributes, a functional form for the iso-preference contours was developed. The following two assumptions were made.

1. The attributes $X$ and $Y$ do not have a linear relationship. This is because a linear relationship implies that the trade-offs are constant regardless of the level of each attribute. It would also imply that the preference for one-unit value change of an attribute is the same as the preference for one-unit value change in another attribute.
2. **Mutual utility independence does not exist** between the attributes $X$ and $Y$.

Mutual utility independence exists when the decision maker’s preferences for one attribute is independent of the level of the other attribute. On a given iso-preference contour the utility independence condition is violated as the DM will be indifferent to two prospects that lie on the same iso-preference contour. For instance, the preference order for the attribute $Y$-number of accidents, without any $X$ ($X=x_0$) is given as $y_0 \succ y_1 \succ y_2 \succ \cdots \succ y_{10}$. However, when the number of fatalities $X$ increases, the preference order for $Y$ is altered and depends on $X$. From fig 4.8b, it can be seen that beyond 8 fatalities ($X \geq 8$), the DM is indifferent to any number of accidents (that is, $y_0 \sim y_1 \sim y_2 \sim \cdots \sim y_{10}$).

The iso-preference contours can be defined as a function of the attributes $X$ and $Y$. For determining a functional preference relationship among the attributes, the generalized form of the Cobb-Douglas production function was used:

$$Z = K x^\alpha y^\beta$$  \hspace{1cm} (4.13)

$K$ is a scaling factor and $\alpha$, $\beta$ are trade-off parameters. $Z$ is a 10x10 matrix that contains the values obtained using the above relation. The values of $Z$ matrix are given in Appendix Table A1.

A set of elements (which form the level set, that is, bounded by the contour lines as shown in the figure 4.8b) from the $Z$ matrix is mapped onto a single preference value.
The preference values have been normalized from 0 to 0.99. The preference value of 1 is assigned to the most preferred prospect (Eq. 4.8). The attributes were grouped into their preference levels by performing a contouring function. With $K$ set to 1, $\alpha$ as 0.95 and $\beta$ as 1.75, the iso-preference contours shown in Fig 4.8b and the corresponding preference levels as shown in Fig 4.8c were obtained. All points on the same iso-preference contour are mapped to the same preference level (Fig 4.8 a).

![Preference for prospects characterized by attributes X and Y](image)

**Fig 4.8a** Preference for prospects characterized by attributes X and Y

Attribute dominance was assumed to exist and this is reflected in the shape of the iso-preference contours in Fig 4.8a. Attribute dominance exists when an attribute exceeds or falls a certain minimum or maximum level, and then the whole preference function is
minimum or maximum, respectively. In this case, the attribute $X$ - Number of fatalities is the dominant attribute. From Fig 4.8a and 4.8b, it can be seen that beyond 8 fatalities, the utility becomes 0. This implies that, when the number of fatalities exceeds 8, the number of accidents is irrelevant and all prospects with $X > 8$ will be preferred equally even if the number of accidents is a minimum for any of the prospects considered.

Fig 4.8b Isopreference Contours of attributes X and Y

4.4.2 Multi-attribute Utility Analysis

When the preference function is a grounded (attribute dominant) preference function, a one-dimensional utility assessment can be used to assess the multi-attribute utility $U (x, y)$ (Abbas 2018). Let $A$ and $B$ be two prospects characterized by their attributes
\((X_a, Y_a)\) and \((X_b, Y_b)\), respectively, as shown in Fig 4.8b. All points on the same iso-preference contour must have the same preference and thus the same utility, as they have the same order ranking (Abbas 2018). Both \(A\) and \(B\) lie on the same iso-preference contour surface and therefore belong to the same preference level of 0.223 (Figure 4.8b).

To assess the utility of prospect \(A\), i.e., the prospect of receiving 3 fatalities and 7 accidents, the following steps were performed.

- Identify a prospect on the same iso-preferential surface at \(Y = y_{\text{min}}\). From Fig 4.8b it can be seen that prospect \(B\) represents the outcome of 1 accident with 6 fatalities.

- Assess the utility function along the curve \(U(x, y_{\text{min}})\), where \(y_{\text{min}}\) represents the prospects of having 1 accident. This is because a fatality cannot occur when there are 0 accidents. This step has already been performed and the one-dimensional utility assessment given in Table 4.4.

From Table 4.4 the utility value of 6 fatalities was elicited as 0.15. Therefore, the utility of prospect \(B\) is \(U(B) = U(6, 1) = 0.15\). Since prospect \(A\) lies on the same preference level as prospect \(B\), it has the same utility as \(B\). Therefore, \(U(A) = U(B)\) and \(U(A) = U(3, 7) = 0.15\). Thus, the utility of a prospect resulting in 3 fatalities and 7 accidents is the same as the utility of a prospect resulting in 6 fatalities and 1 accident and has a utility value of 0.15.
Using a similar process, it is possible to assess the multi-attribute utility of any point (prospect) on the whole domain. The surface of the iso-preference contours projected on the XY plane is shown in Fig 4.8c.

Fig 4.8c. Iso-preference contours projected on XY plane

4.5 Approach 3: Perform Trade-Space Analysis

When safety is defined by two or more objectives but a function to combine them is not available and/or is infeasible, then a Pareto analysis can be performed. This method allows the DM to understand all tradeoffs between the consequences of a decision with respect to all objectives. The Pareto analysis is particularly useful when there is no a priori information regarding the preferences and/or the criteria specified by the DM
(Ngatchou, Zarei et al. 2005). Additionally, the Pareto analysis allows more visualization of the consequences of his/her choices.

Table 4.6 provides the ‘Expected Fatalities’ and ‘Expected Accidents’ for ten notional unmanned aircraft, A through J. The ‘Expected Accidents’ and ‘Expected Fatalities’ values are synthetic and represent probability assessments of fatalities and accidents for each of these aircraft. The ‘Expected Fatalities’ for a particular aircraft are given by the sum of the product of the number of fatalities and its corresponding probability of occurrence. Similarly, the ‘Expected Accidents’ for a particular aircraft is given by the sum of the product of the number of accidents and its corresponding probability of occurrence. Source synthetic data are given in the Appendix Table A2.1, A2.2. The probability assessments are assumed to span equal amount of flight hours for all 10 aircraft. The option space consists of the ‘Expected Fatalities’ and ‘Expected Accidents’ for aircraft A through J presented in Fig 4.9.

Table 4.6 ‘Expected Fatalities’ and ‘Expected Accidents’ for 10 Aircraft Types

<table>
<thead>
<tr>
<th>Aircraft Type</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected [Fatalities]</td>
<td>1.189</td>
<td>1.605</td>
<td>0.189</td>
<td>1.357</td>
<td>1.624</td>
<td>1.004</td>
<td>2.846</td>
<td>2.963</td>
<td>1.085</td>
<td>0.3357</td>
</tr>
<tr>
<td>Expected [Accidents]</td>
<td>0.8566</td>
<td>0.8</td>
<td>2.6462</td>
<td>1.7104</td>
<td>0.534</td>
<td>0.7537</td>
<td>1.3181</td>
<td>0.84</td>
<td>1.5702</td>
<td>1.502</td>
</tr>
</tbody>
</table>
In this example, as previously, less ‘Expected Fatalities’ are preferred over more, and less ‘Expected Accidents’ are preferred over more. However, the mutual preference between fatalities and accidents is assumed unknown. In this case, the Pareto frontier is given by those solutions that are not dominated. The set is given by aircraft C, J, F, and E. Note that, for example, although aircraft C has a high number of ‘Expected Accidents’, no other solution can beat it in terms of ‘Expected Fatalities’. The contrary occurs with aircraft E, which has a high number of ‘Expected Fatalities’, but not solution yields lower ‘Expected accidents’. Dominated solutions, those not in the Pareto front, are those for which a better alternative can be found. For example, if someone were to choose solution I, it is easy to see that aircraft J would provide a lower number of ‘Expected Accidents’ and lower number of ‘Expected Fatalities’, so there would no reason to prefer J over I. Similar reasoning can be applied to aircraft D, A, B, H, and G in this example.
From a preference standpoint, a decision-maker should be indifferent between the four aircraft with the information at hand. Visualization is expected to foster discussions that enable eventually choosing one of the aircraft, as an order between the outcomes of those specific solutions might be established (Fitzgerald and Ross 2015).

4.6 Expert Assessment Aggregation

In this section problems pertaining to expert aggregation in safety risk assessment are discussed. The solutions discussed in chapter 2 were applied to safety risk assessment aggregation and the implications of the proposed techniques are elaborated. In the next section the illustration of the discursive dilemma in safety risk assessment is shown.

4.6.1 Judgement Aggregation in Safety Risk Assessment: The Discursive Dilemma

Consider a safety risk assessment required to be performed by a team of three experts. The judgement of all three experts is assumed to be equally trustable and relevant for the decision. There is a debate between the severity and likelihood of the hazard under consideration. The hazard will be classified as High Risk if the severity is judged as level 3 (Major) and the likelihood of occurrence is judged as A (Frequent) (Fig 4.1). The SMEs are participating in a group safety risk assessment task and are asked their judgement for severity and likelihood of a hazard. The SMEs are required to provide their judgement to the following questions.

- Is the severity of the Hazard 3 (Major)?
- Is the likelihood of the Hazard A (Frequent)?
For the purpose of the argument, it is assumed that the severity of the hazard can have a maximum severity level of 3 (Major).

It is assumed that each SMEs judgement set (severity, likelihood and safety risk) satisfies the three constraints of completeness, consistency and deductive closure (List and Pettit 2002). An SMEs personal judgement set is said to be complete, consistent and deductively closed if, for all propositions proposed by the SME, the final judgement is available in the universal domain set (High, Medium, Low Risk), conforms to a logical deduction and the logic remains true for the combinations of the propositions. The goal is to find an aggregation technique that conforms to the minimum requirements of completeness, consistency and deductive cogency (List and Pettit 2002). The SMEs judgement follows the logic:

\[(\text{Severity (Major 3)} \land \text{Likelihood (Frequent A)}) \rightarrow \text{Safety Risk (High Risk)}\]

Table 4.7 Discursive Dilemma in Safety Risk Assessment

<table>
<thead>
<tr>
<th>SME</th>
<th>Severity</th>
<th>Likelihood</th>
<th>Safety Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Major (3)</td>
<td>Frequent (A)</td>
<td>High Risk</td>
</tr>
<tr>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Majority</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

If the DM proceeds with a majority rule, the requirements of deductive cogency and consistency are violated. The majority asserts that the severity of the hazard is Major, and
that the likelihood of occurrence is frequent. Yet, the majority rules out the proposition that the safety risk of the hazard is High.

The impossibility result also shows that such problems of deductive cogency and consistency are not only confined to a majority procedure but also persistent in other aggregation procedures (List and Puppe 2009, Hartmann, Pigozzi et al. 2010). The other aggregation procedures that might be affected by the impossibility results include unanimitarianism (virtue of pertaining to unanimity), minoritarianism and other procedures that have conditional majoritarianism such as two-thirds (List and Pettit 2002). This further indicates that if the group follows some behavioral aggregation or even a structured aggregation process (van Steen 1992) of coming to a group consensus, then they might also fall into the problems of the discursive dilemma.

4.6.2 Implications of the dilemma
The FAA and the U.S. Navy both stress the need to use proper documentation. Once a safety risk has been assessed and documented, then the decision process is solely based on the documented value of the safety risk (Navy-OPNAVINST3500.39C 2010, FAA-Order8040.4B 2017). If the hazard has been determined as high risk, the decision process is assigned to higher authoritarian members of the organization. If the safety risk has been assessed as medium or low, then the decision authority or advisors responsible for providing instruction for the mitigation and control activities of the hazard are members of lower authoritative ranks. Hence, there is a risk that the rightful authorities might not
be called to act on the hazard that has been incorrectly judged as *Medium, Low, or High* as a result of the dilemma.

The implications of the discursive dilemma not only stop at the organizational process level but might have far reaching implications that lead to inconsistent decisions. For instance, for the same judgement sets, a hazard could be classified both as *High* risk and *Medium* or *Low* risk at different times, causing mitigation strategies to vary each time. When the mitigation and control strategies have not been implemented rightfully, then a high-risk hazard could be neglected and might result in a catastrophic situation whose investigation cannot be traced back to a rationale.

Another plausible negative implication could be that, when undesirable outcomes emerge as a consequence of the decision based on the safety risk assessment, the aggregation procedure might not be questioned but rather the SMEs judgement (prediction accuracy) might be deemed unreliable.

### 4.7 Evaluation of Potential Solutions to the Discursive Dilemma

#### 4.7.1 Practicing CBP and PBP

If CBP is being practiced, then the individual SMEs calculate the safety risk based on their personal assessment of severity and likelihood. Their conclusion is then aggregated. The result of CBP with respect to the premises of the experts is presented in Table 4.8. The result shows that the majority would decide that the safety risk is **NOT** High Risk.
When PBP is practiced, the group’s assessments of the severity and likelihood is first aggregated and then a conclusion is drawn from the aggregation. From Table 4.9, the Severity level 3 is voted as Yes by the majority, and the likelihood, Frequent A, is also voted as Yes by the majority. Therefore, the Safety Risk is judged as **High Risk** based on the logical relationship

\[(\text{Severity (Major 3)} \land \text{Likelihood (Frequent A)}) \rightarrow \text{Safety Risk (High risk)}\]

From this it can be seen that the same experts having the same judgement can potentially end up with different assessments depending on the questions asked. This indicates
inconsistency and retraces back to the problems in expert elicitation (Goodwin and Wright 2014). A normative aggregation method, however, should yield the same assessment, regardless of the method of elicitation and should abide by the logic used to derive the judgement.

One important result of Hartmann, Pigozzi et al. (2010) is that, in a voting procedure, the reasons should carry more weight than the conclusion. This implies that collective rationality should be maintained and carries more significance. Therefore, we can choose to forego adherence to individual responsiveness. In Safety Risk Assessment, following PBP leads to taking the ‘safer’ route, since the relationship between the conclusion and the premises is essentially a logical “And” operation. In a worst-case scenario, hence, PBP leads to judge a safety risk as higher than what was intended by the individual SMEs.

4.7.2 Practicing modus tollens instead of modus ponens

Although practicing *modus tollens* could be a possible technique to avoid the dilemma, it could have negative implications in Safety Risk Assessment. If modus tollens is practiced, then the SMEs will be asked their judgement on either the severity or the likelihood and their opinion on the Safety Risk. Then the unquestioned variable (either the severity or the likelihood) can be inferred from the relation. This method maintains deductive cogency and collective reasoning and is a strategy to evade the dilemma.
Consider the following example. The experts are asked if the safety risk of the hazard is *High*. If their answer is affirmative, then the severity of the hazard is *Major* (3) and the frequency of the hazard is *Frequent* (A). If one parameter is fixed, then the modus tollens method can be used to identify an inconsistency in the judgement of an SME. Consider the judgement set of SME 3 in Table 1.5. The SME thinks that the safety risk posed by the hazard is severe but does not seem to accept that the frequency of the hazard is Frequent. Such problems are attributed to inconsistencies in expert elicitation and behavioral aggregation, where the judges may violate principles of deductive cogency without realizing its implications.

**Table 4.10 Modus Tollens**

<table>
<thead>
<tr>
<th>SME</th>
<th>Severity</th>
<th>Likelihood</th>
<th>Safety Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>Major</em> (3)</td>
<td><em>Frequent</em> (A)</td>
<td><em>High Risk</em></td>
</tr>
<tr>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>No?</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### 4.7.3 Other Strategies to Evade the Dilemma

List and Pettit (2002) present 4 strategies where the aggregation procedure can yield “collectivized reason”. We apply those strategies to our problem and analyze the outcomes.

1. **Using a Convergence Strategy**

By practicing interpersonal deliberations or by other methods, if the views of the judges are made to converge, then the conclusion of the collective majority will be complete,
consistent and deductively closed. By doing so we are relaxing the universal domain. This suggests adopting a behavioral technique where the group comes to a consensus on the safety risk assessment.

It should be noted that, while this strategy can be useful to avoid the paradox, it may do so at the expense of becoming unreliable. When interactions occur among the experts, it might be difficult to distinguish if convergence occurs because the knowledge base of the experts has now unified or because the group is biased towards the same opinion. It is also difficult to acknowledge group level consequences that emerge as a result of individual interaction systems (Davis 1992). The group interactions could also give rise to a number of cognitive biases (Das and Teng 1999) that might go undetected.

**ii. Using an Authoritative Strategy**

In the above majority voting, each expert was considered to have equal weight and all their assessments were treated as equally important. However, if modifying the group structure is permitted so that only one members input is taken into consideration, then, it is easy to avoid the paradox. In this case, the rule of anonymity is relaxed. This however brings back the problems of assigning proper weights to the experts. Furthermore, considering the judgements of only a limited number of experts will put into question why the other experts were included in the assessment in the first place. Ignoring those experts may lead to selection biases (Norton, Sommers et al. 2007), where only information that the DM wants to hear is used to make the decision.
iii. Using a Priority Strategy

In this strategy, the group must decide on a set of propositions to be given higher priority over the others. The group’s decision process of the other set of propositions differs from the prioritized set and is determined on the overall decision of the prioritized set of positions. If a conclusive judgement set is contradictory and/or does not belong to the prioritized judgement set, then the group eliminates the judgement set that has a lower priority, thus eliminating its impact on the final judgment. By following a priority strategy, the rule of systematicity is relaxed. The implication of relaxing the rule of systematicity can be that the logic used to arrive at the aggregated judgement might be inconsistent. Although the dilemma might be evaded, the rationale behind the aggregated judgement might be incorrect.

iv. Using a Special-Support Strategy

In this strategy, a proposition set proposed by each expert must be endorsed or receive “special support” by the majority of other experts. It can be called as a “majority of the majorities”. Each expert presents his/her judgment set to the remainder of the group and asks for endorsements. The judgement set that has the highest endorsements or support by the ‘*super-majority*’ is concluded as the final collective judgement. By using a Special-Support strategy, the rule of completeness is relaxed. If the expert receiving the ‘special support’ does not have a logical judgement, then again rationale behind the safety risk assessment could be flawed. Further, this strategy also suffers from a potential bias of ‘group think’.
A summary of the implications of using each of these four strategies in safety risk assessment is given in Table 4.11

Table 4.11 Aggregation Strategies that yields “Collectivized Reason”

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convergence Strategy</td>
<td>Unreliable method: gives rise to potential group biases that might go undetected.</td>
</tr>
<tr>
<td>Authoritative Strategy</td>
<td>Assigning weights or determining the authoritarian poses challenging; might lead to selection bias.</td>
</tr>
<tr>
<td>Priority Strategy</td>
<td>Logic between premise and conclusion might become deterred.</td>
</tr>
<tr>
<td>Special Support Strategy</td>
<td>Rationale behind the safety risk assessment might be illogical; leads to potential group bias.</td>
</tr>
</tbody>
</table>

4.8 Mathematical Aggregation

In this section the mathematical aggregation techniques discussed in Chapter 2 are applied to safety risk assessment and the results are discussed.

4.8.1 Arithmetic and Geometric Averaging for Aggregating Severity Rankings

In the FAA’s SRMP, the severity rankings of the hazards are mapped onto an arbitrary scale of 1 – 5 with 5 indicating minimal to no effect and 1 being highly catastrophic in nature (FAA-Order8040.4B 2017). With these severity rankings as references, a panel of 5 experts provide their assessments on 3 potential hazards X, Y and Z to an external DM. The DM performs arithmetic (AM) and geometric (GM) averaging of the assessments. Table 4.12 shows the assessments from the experts on hazards X, Y and Z.
Table 4.12 Severity Assessments for Hazards X, Y and Z by 5 experts

<table>
<thead>
<tr>
<th>Expert</th>
<th>Hazard X</th>
<th>Hazard Y</th>
<th>Hazard Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 1</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Expert 2</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Expert 3</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Expert 4</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Expert 5</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>AM</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>GM</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

It can be seen that the choice of averaging could yield different results. This is mainly due to the mathematical calculation behind the averaging method. AM is the sum of the assessments divided by the number of assessments and GM is the $n^{th}$ root of the product of the assessments, where $n$ is the number of assessments. In the geometric averaging, the effect of ‘outliers’ is dampened. However, since the possible set of assessments here lie only in the range of 1 to 5, the characteristic of dampening outliers by the GM might not make a critical impact. One advantage of such averaging methods is its computational simplicity though.

In aggregating severity rankings, when the assessments of the experts vary between extreme points, then applying either AM or GM discounts the assessment of the expert and might lead to regretful decision outcomes. For instance, in the assessments of Hazard X and Hazard Z (Table 4.12), Expert 3 has assessed the severity of that hazard to be catastrophic. However, when the assessments of all the experts was averaged, the
aggregated assessment indicated that the hazard is not as ‘severe’ as anticipated by Expert 3. In a situation where Expert 3 has the highest prediction accuracy, then the averaged aggregated assessment has discounted the input of Expert 3 and has yielded a lower severity level. Thus, leading to improper safety risk assessment and thereby to inadequate mitigation and control strategies required for the hazard.

A weighted averaging scheme may be used to overcome such situations. However, those schemes are also subjected to the same problems of determining the weights of the individual experts described in Chapter 2. Therefore, they are not effective solutions to aggregate expert assessments for severity rankings.

4.8.2 Belief Aggregation – Safety Risk Assessment

The FAA defines likelihood as “the estimated probability or frequency, in quantitative or qualitative terms, of a hazard’s effect or outcome”. (FAA-Order8040.6 2019) Table 4.13 provides the likelihood definitions followed by the FAA based on which the experts assert their assessments for a determining the safety risk of a given hazard.
Table 4.13 Likelihood Definitions (Source: Appendix C Table C2. (FAA-Order8040.6 2019))

<table>
<thead>
<tr>
<th>Frequent A</th>
<th>Expected to occur routinely</th>
<th>Expected to occur more than 100 times per year (or more than approximately 10 times a month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probable B</td>
<td>Expected to occur often</td>
<td>Expected to occur between 10 and 100 times per year (or approximately 1-10 times a month)</td>
</tr>
<tr>
<td>Remote C</td>
<td>Expected to occur infrequently</td>
<td>Expected to occur one time every 1 month to 1 year</td>
</tr>
<tr>
<td>Extremely Remote D</td>
<td>Expected to occur rarely</td>
<td>Expected to occur one time every 1 to 10 years</td>
</tr>
<tr>
<td>Extremely Improbable E</td>
<td>Unlikely to occur, but not impossible</td>
<td>Expected to occur less than one time every 10 years</td>
</tr>
</tbody>
</table>

It should be noted that there is a possibility that the experts are not in complete agreement with the definitions of the hazard likelihood ranges. The degree of disagreement between the definitions of the uncertain elements should be addressed as an integral part of the aggregation problem (Bonduelle 1988). Even assuming that the experts’ agree on the definitions of the likelihood, it is still possible for them to disagree about the probability of hazard occurrence (Clemen and Winkler 1999). The terms probability and likelihood are quite different, yet it is possible to use the terms interchangeably in the context of safety risk assessment. The method of expert elicitation is to be scrutinized to overcome such problems (Walls and Quigley 2001).

When two experts assert the likelihood of a hazard being probable, their probability (belief) distribution could be quite different from each other’s. Here, the belief distributions represent the likelihood of the hazards occurring at a given mean rate over
the period of 1 year. For instance, both belief distributions could follow a Poisson process yet can have different mean rates of failure. This implies that their probability of the number of times a hazard occurs varies with the distribution. For example, Figure 4.10 shows a notional situation in which both experts A and B have asserted ‘probable’ for their respective belief of the mean frequency of occurrence. For a hazard occurring less than or equal to 160 times in one-year, expert A has a belief probability of 0.99 whereas expert B has a belief probability of 0.8.

![Fig 4.10 Probability distribution of Experts A and B](image)

### 4.8.3 Belief Aggregation for Likelihood of Hazard Occurrence

This section provides an illustration of linear and logarithmic opinion pooling of the judgement of 5 experts who are participating in a safety risk assessment of an unmanned aircraft. The scenario description of the hazard has been leveraged from Appendix A of the FAA’s UAS Safety Risk Management Policy (FAA-Order8040.6 2019).
The Hazard has been called as having ‘A technical issue with UAS’ and there are 5 potential reasons for the hazard to materialize. Each reason is associated with a different sub-system/component and the relevant expert is asked for his/her judgement on the on the likelihood of failure of that sub-system/component. Table 4.14 provides the causes for the Hazard, the expert involved and the experts’ judgement on the likelihood.

<table>
<thead>
<tr>
<th>Hazard</th>
<th>Causes</th>
<th>Expert Assigned</th>
<th>Expert Judgement on Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical Issue with UAS</td>
<td>Motor Failure</td>
<td>Expert 1</td>
<td>Frequent</td>
</tr>
<tr>
<td></td>
<td>Software Failure</td>
<td>Expert 2</td>
<td>Frequent</td>
</tr>
<tr>
<td></td>
<td>Avionics Failure</td>
<td>Expert 3</td>
<td>Probable</td>
</tr>
<tr>
<td></td>
<td>GPS Failure</td>
<td>Expert 4</td>
<td>Remote</td>
</tr>
<tr>
<td></td>
<td>Lost Link</td>
<td>Expert 5</td>
<td>Remote</td>
</tr>
</tbody>
</table>

The belief distributions of each of the experts is assumed to follow a Poisson process, each with a unique mean rate of equipment failure. A Poisson distribution expresses the probability of a given number of independent events occurring over a fixed time interval having a constant mean rate. Given that the event of a sub-subsystem failing can be represented as discrete and are considered to be independent, a Poisson distribution was used to represent the belief of the experts.

If $X$ is a discrete random variable and $k$ is the number of times a given cause of failure (Table 4.13) occurs per year ($k = 1, 2, 3, ... n$) with a mean rate of failure $\lambda$, then the probability distribution is given by equation 4.14.
\[ P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (4.14) \]

From Table 4.14, consider the notional situation in which the 5 individual experts have asserted their likelihood estimations for their respective component failures. Table 4.15 represents the mean rate of failure that the SMEs associate with the asserted estimations. Figure 4.11 represents the respective belief distributions based on the mean rate gives in table 4.15. Two methods of aggregating these belief distributions were implemented – Linear pooling and Logarithmic Pooling.

<table>
<thead>
<tr>
<th>SME</th>
<th>Qualitative Likelihood Metric Associated with their Belief</th>
<th>Mean Rate ((\lambda))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 1</td>
<td>Frequent</td>
<td>50</td>
</tr>
<tr>
<td>Expert 2</td>
<td>Frequent</td>
<td>70</td>
</tr>
<tr>
<td>Expert 3</td>
<td>Probable</td>
<td>35</td>
</tr>
<tr>
<td>Expert 4</td>
<td>Remote</td>
<td>10</td>
</tr>
<tr>
<td>Expert 5</td>
<td>Remote</td>
<td>2.5</td>
</tr>
</tbody>
</table>
4.8.4 Linear Pooling

Linear pooling of the distributions is the weighted linear combination of the 5 belief distributions of the 5 experts (Fig 4.11). Each expert’s belief distribution was assigned an equal weight of 0.2. The linear pooling is expressed as,

\[ \pi_{linear}(k) = \sum_{i=1}^{5} w_i \pi_i(k) \]  \hspace{1cm} (4.15)

Here

\[ i = 1, 2, 3, 4, 5 \]  \hspace{1cm} and \hspace{1cm} \[ w_1 + w_2 + w_3 + w_4 + w_5 = 1 \]
The averaged distribution after performing linear pooling of the individual belief distributions is shown in Fig 4.12. The mean rate of the averaged distribution was found to be \( \lambda_{\text{avg}} \) (linear) = 34.

At the mean rate of 34, the cumulative probability is 0.54 for the averaged distribution (ref. Fig. 4.13).
4.8.5 Logarithmic Pooling

The logarithmic pooling of the 5 belief distributions of the 5 experts based on the equation 4.16 is given in figure 4.14 and the CDF of the distributions is given in Fig 4.17.

\[ \pi_{\log}(k) = \prod_{i=1}^{n} (\pi_i(k))^{w_i} \]  

(4.16)

For logarithmic pooling, the mean rate of the averaged distributions was obtained as 20; 
\[ \lambda_{\text{avg}}(\log) = 20. \]
This implies that the likelihood of occurrence is approximately 50% of the number of times a cause of failure occurs, that is less than or equal to 34 for linear pooling and less than or equal to 24 for logarithmic pooling mechanisms. Therefore, the method of pooling will have a significant impact on the determination of the likelihood of the hazards. However, the focus of this research is not to debate between the advantages and disadvantages between the two methods, but rather to illustrate that there can be significant differences in determining the safety risk level of the hazard based on the chosen aggregation technique.

Fig 4.14 Logarithmic Pooling
4.8.6 Using Likelihood Assessment as a Judgment rather than a Belief

When the probability distributions (belief) of the experts are made available, then aggregating them using linear and logarithmic pooling is reasonable. However, when an expert asserts a likelihood value, he/she is confined to the definitions of likelihood presented in table 4.12 and can be viewed as a judgement rather than a ‘belief’. In such scenarios, coming to an aggregated conclusive judgment i.e. labeling a hazard as ‘Frequent’, ‘Probable’, ‘Remote’ etc. based on an aggregated distribution calls for sensitivity analysis methods (Clemen and Winkler 1999) employed in domains such as reliability engineering.
CHAPTER 5

Conclusion

This chapter summarizes the contributions of this research. The answers for the two research questions RQ1 and RQ2 are recapitulated. The recommendations for safety risk assessment practices based on the results obtained in chapter 4 is discussed. Finally, the limitations of this research and future work is presented.

5.1 Contributions, Conclusions and Recommendations

This thesis has contributed to the safety risk assessment of SCSs in three aspects. First it scrutinized the current practices of safety risk assessment from a decision analysis standpoint. Next, it evaluated the use of decision analysis for safety risk assessment in safety decision making. Third, it reviewed the expert aggregation mechanisms that preserve the premise and judgement consistency in safety risk assessment. The subsequent paragraphs provide a summary of the conclusions and recommendations of this research.

The current practices of using a risk matrix for safety risk assessment leads to inconsistencies in the safety decision process as it is subjective to the nature of participation of the SMEs. Further, the elements of the risk matrix, both severity and likelihood are defined using high-level qualitative statements which are prone to subjective interpretations of the SMEs and the DM. Table 5.1 presents the inconsistencies of current methods of safety risk assessment that was identified in this research.
Table 5.1 Inconsistencies of Current Methods of Safety Risk Assessment

<table>
<thead>
<tr>
<th>Safety Risk Assessment Practice</th>
<th>Implication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safety Risk Matrix</td>
<td>Ordering of the severity rankings do not account for the meaning of the severity classes.</td>
</tr>
<tr>
<td></td>
<td>Severity definitions encapsulate many safety metrics into a singular definition mapped to an arbitrary score.</td>
</tr>
<tr>
<td></td>
<td>Likelihood definitions fail to reflect the beliefs of SMEs</td>
</tr>
<tr>
<td>Expert Aggregation</td>
<td>No prescribed method to aggregate expertise, potentially giving rise to inconsistent assessment of the safety risk as a result of the discursive dilemma.</td>
</tr>
<tr>
<td></td>
<td>- Behavioral methods lead to problems such as group think, selection bias and expert selection/calibration.</td>
</tr>
<tr>
<td></td>
<td>Current method of safety risk assessment does not serve as a good platform to implement mathematical aggregation of expertise.</td>
</tr>
<tr>
<td></td>
<td>- Mathematical aggregation of arbitrary scores of severity rankings leads to discounting expert assessments.</td>
</tr>
<tr>
<td></td>
<td>- Mathematical Aggregation of judgements on likelihood leads to inferential inconsistencies.</td>
</tr>
</tbody>
</table>
Upon analysis of the elements of the safety risk assessment process, the problems associated with the severity was identified. The definitions for severity include multiple factors such as accidents, fatalities, property damage, equipment damage etc. Encapsulating this many metrics into a single qualitative definition is highly prone to err. This inconsistency was proved by applying utility theory to the severity classes of the elements of the risk matrix. It was observed that the preference order of the elements differed from the presently used preference order classification. To overcome this inconsistency and meaningfully combine multiple safety metrics in the safety decision process, 3 approaches were proposed (RQ1).

i. *Use a single metric and apply Utility Theory.* This method allows capturing the preferences of the DM with respect to a single metric. This method is advantageous for its simplicity. The risk attitude of the DM can also be assessed. However, it demands a quantified metric in order to assign utility values.

ii. *Combine two metrics using Multi-attribute Utility Theory.* Here two safety metrics are combined, and a utility value is calculated for the coupled metric. The preference function developed captures the relative trade-off preferences of the DM for the two different metrics. This method could be advantageous to use instead of high-level severity definitions. However, the analysis and development of preference/value functions becomes computationally
complex as the number of metrics increases. Especially when metrics have a level of dependency associated with them, eliciting the trade-off parameters will pose challenges.

iii. *Perform a Pareto Analysis.* This method preserves the inferential criticality of the individual safety metrics. This method is computationally simple and also serves as a visual tool that can enable meaningful discussions regarding safety trade-off parameters.

With respect to aggregation of expertise in safety risk assessment (RQ2), it was deduced that expert judgement was needed for both hazard severity and likelihood determination. The plausibility of the discursive dilemma in safety risk assessment aggregation was demonstrated. Potential solutions were tested and their implications in safety risk assessment was discussed. With respect to safety risk assessment, it was found that, if a judgement aggregation must take place, then a ‘*better*’ process to evade the dilemma would be PBP (Premise based procedure). However, all solutions explored had drawbacks.

Using the current definitions of severity, which are abstract and qualitative, and using them to perform a mathematical aggregation of the severity judgments leads to problems such as damping critical safety risk assessments and discounting minority assessments (Cassandra Syndrome). These problems reassert the use of utility theory as a potential solution for assessing severity of a hazard.
With respect to likelihood, the problem of treating likelihood assessments as judgements was illustrated. The likelihood defined in probability ranges, is inadequate to capture the belief of an expert. Therefore, for likelihood, the belief distributions of the SMEs should be elicited and pooled rather than aggregating their judgements. The discussion of the most appropriate method of belief aggregation (linear/ logarithmic/ Bayesian) for safety risk assessment is proposed for future work.

5.2 Research Limitations

There are two main limitations of this research. First, the arguments developed in this research were based on notional cases and the data used was synthetic. Yet the assumptions made captured reasonable behavior of a DM with respect to safety decision making. Furthermore, other external factors such as mission cost, timelines and performance requirements that could impact the DM’s choices were not accounted for.

Second, safety metrics considered to illustrate the use of Utility Theory had a small range of values with a maximum of 10 fatalities and 10 accidents. However, in reality a larger dataset could be the considered. In such cases, eliciting utilities can be highly difficult and utility theory might not be a feasible approach. Additionally, the problems associated with different methods of utility elicitation were not explored. Thereby, their impact on safety risk assessment was not evaluated.
5.3 Future Work

The current methods of performing safety risk assessment need to be improved. The benefits of Decision Analysis for safety decisions in SCSs should be explored further. An empirical analysis could be used to emphasize the advantages and disadvantages of the utility theory approach to combine safety metrics. Also, methods to reduce reliance on human expertise for safety risk assessment could be developed.

In judgement aggregation, recent developments make use of approximation methods employed in computer science and artificial intelligence can be explored. Distance based procedures – such as the use of a Majority Fusion Operator (MFO) (Pigozzi 2006), which yields an aggregated judgement that is closest to the group average and neural net methods (Mak, Bui et al. 1996) could be applied to safety risk assessment and its efficacy can be evaluated. For belief aggregation, Bayesian Methods should be explored for safety risk assessment and more emphasis on expert elicitation and calibration should be given.
References


Appendix A

Table A1 Z - Matrix Values (After Normalization)

<table>
<thead>
<tr>
<th></th>
<th>0.998009</th>
<th>0.829963</th>
<th>0.67537</th>
<th>0.534633</th>
<th>0.408224</th>
<th>0.29671</th>
<th>0.200789</th>
<th>0.121366</th>
<th>0.059695</th>
<th>0.017747</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.902952</td>
<td>0.750912</td>
<td>0.611044</td>
<td>0.483711</td>
<td>0.369343</td>
<td>0.268449</td>
<td>0.181664</td>
<td>0.109806</td>
<td>0.054009</td>
<td>0.016057</td>
<td></td>
</tr>
<tr>
<td>0.807365</td>
<td>0.67142</td>
<td>0.546358</td>
<td>0.432505</td>
<td>0.330244</td>
<td>0.240031</td>
<td>0.162433</td>
<td>0.098182</td>
<td>0.048292</td>
<td>0.014357</td>
<td></td>
</tr>
<tr>
<td>0.711177</td>
<td>0.591428</td>
<td>0.481266</td>
<td>0.380977</td>
<td>0.290899</td>
<td>0.211434</td>
<td>0.143081</td>
<td>0.086485</td>
<td>0.042538</td>
<td>0.012647</td>
<td></td>
</tr>
<tr>
<td>0.614296</td>
<td>0.510861</td>
<td>0.415705</td>
<td>0.329079</td>
<td>0.251271</td>
<td>0.182631</td>
<td>0.12359</td>
<td>0.074703</td>
<td>0.036743</td>
<td>0.010924</td>
<td></td>
</tr>
<tr>
<td>0.516602</td>
<td>0.429616</td>
<td>0.349593</td>
<td>0.276743</td>
<td>0.21131</td>
<td>0.153587</td>
<td>0.103935</td>
<td>0.062823</td>
<td>0.024997</td>
<td>0.007432</td>
<td></td>
</tr>
<tr>
<td>0.417918</td>
<td>0.347549</td>
<td>0.282813</td>
<td>0.223879</td>
<td>0.170945</td>
<td>0.124248</td>
<td>0.084081</td>
<td>0.050822</td>
<td>0.024997</td>
<td>0.007432</td>
<td></td>
</tr>
<tr>
<td>0.31798</td>
<td>0.264438</td>
<td>0.215183</td>
<td>0.170342</td>
<td>0.13066</td>
<td>0.094536</td>
<td>0.063974</td>
<td>0.038669</td>
<td>0.01902</td>
<td>0.005655</td>
<td></td>
</tr>
<tr>
<td>0.216328</td>
<td>0.179903</td>
<td>0.146393</td>
<td>0.115887</td>
<td>0.088487</td>
<td>0.064315</td>
<td>0.043523</td>
<td>0.026307</td>
<td>0.012939</td>
<td>0.003847</td>
<td></td>
</tr>
<tr>
<td>0.111978</td>
<td>0.093123</td>
<td>0.075778</td>
<td>0.059987</td>
<td>0.045804</td>
<td>0.033291</td>
<td>0.022529</td>
<td>0.013617</td>
<td>0.006698</td>
<td>0.001991</td>
<td></td>
</tr>
</tbody>
</table>

Table A2.1 Fatality Probabilities for Aircraft A, B, C, ..., J

<table>
<thead>
<tr>
<th>Number of Fatalities</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.65</td>
<td>0.5</td>
<td>0.87</td>
<td>0.5</td>
<td>0.45</td>
<td>0.5</td>
<td>0.5</td>
<td>0.6</td>
<td>0.68</td>
<td>0.85</td>
</tr>
<tr>
<td>1</td>
<td>0.071</td>
<td>0.171</td>
<td>0.112</td>
<td>0.2</td>
<td>0.171</td>
<td>0.213</td>
<td>0.04</td>
<td>0.001</td>
<td>0.071</td>
<td>0.052</td>
</tr>
<tr>
<td>2</td>
<td>0.068</td>
<td>0.08</td>
<td>0.006</td>
<td>0.09</td>
<td>0.13</td>
<td>0.137</td>
<td>0.04</td>
<td>0.031</td>
<td>0.06</td>
<td>0.041</td>
</tr>
<tr>
<td>3</td>
<td>0.06</td>
<td>0.06</td>
<td>0.004</td>
<td>0.06</td>
<td>0.067</td>
<td>0.092</td>
<td>0.051</td>
<td>0.02</td>
<td>0.052</td>
<td>0.04</td>
</tr>
<tr>
<td>4</td>
<td>0.052</td>
<td>0.052</td>
<td>0.002</td>
<td>0.052</td>
<td>0.059</td>
<td>0.052</td>
<td>0.042</td>
<td>0.02</td>
<td>0.041</td>
<td>0.004</td>
</tr>
<tr>
<td>5</td>
<td>0.041</td>
<td>0.041</td>
<td>0.001</td>
<td>0.04</td>
<td>0.047</td>
<td>0.004</td>
<td>0.047</td>
<td>0.028</td>
<td>0.039</td>
<td>0.0125</td>
</tr>
<tr>
<td>6</td>
<td>0.03</td>
<td>0.031</td>
<td>0.001</td>
<td>0.03</td>
<td>0.035</td>
<td>0.001</td>
<td>0.042</td>
<td>0.23</td>
<td>0.03</td>
<td>0.0003</td>
</tr>
<tr>
<td>7</td>
<td>0.02</td>
<td>0.031</td>
<td>0.001</td>
<td>0.02</td>
<td>0.029</td>
<td>0.001</td>
<td>0.087</td>
<td>0.03</td>
<td>0.02</td>
<td>0.0002</td>
</tr>
<tr>
<td>8</td>
<td>0.004</td>
<td>0.03</td>
<td>0.001</td>
<td>0.004</td>
<td>0.004</td>
<td>0</td>
<td>0.064</td>
<td>0.02</td>
<td>0.004</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0.003</td>
<td>0.002</td>
<td>0.001</td>
<td>0.003</td>
<td>0.004</td>
<td>0</td>
<td>0.073</td>
<td>0.01</td>
<td>0.003</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0.001</td>
<td>0.002</td>
<td>0.001</td>
<td>0.001</td>
<td>0.004</td>
<td>0</td>
<td>0.014</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E[Fatalities]</td>
<td>1.189</td>
<td>1.605</td>
<td>0.189</td>
<td>1.357</td>
<td>1.624</td>
<td>1.004</td>
<td>2.846</td>
<td>2.963</td>
<td>1.085</td>
<td>0.3357</td>
</tr>
</tbody>
</table>
Table A2.2 Accident Probabilities for Aircraft A, B, C, ..., J

<table>
<thead>
<tr>
<th>Number of Accidents</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.57</td>
<td>0.69</td>
<td>0.7</td>
<td>0.4</td>
<td>0.65</td>
<td>0.75</td>
<td>0.43</td>
<td>0.53</td>
<td>0.45</td>
<td>0.2</td>
</tr>
<tr>
<td>1</td>
<td>0.3064</td>
<td>0.168</td>
<td>0.0001</td>
<td>0.3084</td>
<td>0.272</td>
<td>0.104</td>
<td>0.3</td>
<td>0.30</td>
<td>0.3084</td>
<td>0.369</td>
</tr>
<tr>
<td>2</td>
<td>0.030</td>
<td>0.030</td>
<td>0.0001</td>
<td>0.08</td>
<td>0.032</td>
<td>0.030</td>
<td>0.0358</td>
<td>0.06</td>
<td>0.03</td>
<td>0.29</td>
</tr>
<tr>
<td>3</td>
<td>0.021</td>
<td>0.025</td>
<td>0.0002</td>
<td>0.031</td>
<td>0.015</td>
<td>0.0259</td>
<td>0.1421</td>
<td>0.05</td>
<td>0.025</td>
<td>0.1</td>
</tr>
<tr>
<td>4</td>
<td>0.007</td>
<td>0.007</td>
<td>0.0003</td>
<td>0.007</td>
<td>0.014</td>
<td>0.025</td>
<td>0.0124</td>
<td>0.04</td>
<td>0.017</td>
<td>0.011</td>
</tr>
<tr>
<td>5</td>
<td>0.006</td>
<td>0.04</td>
<td>0.006</td>
<td>0.0546</td>
<td>0.011</td>
<td>0.02</td>
<td>0.032</td>
<td>0.014</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>6</td>
<td>0.052</td>
<td>0.022</td>
<td>0.042</td>
<td>0.052</td>
<td>0.002</td>
<td>0.02</td>
<td>0.025</td>
<td>0.003</td>
<td>0.041</td>
<td>0.005</td>
</tr>
<tr>
<td>7</td>
<td>0.0042</td>
<td>0.01</td>
<td>0.0083</td>
<td>0.042</td>
<td>0.002</td>
<td>0.012</td>
<td>0.0214</td>
<td>0.002</td>
<td>0.0374</td>
<td>0.003</td>
</tr>
<tr>
<td>8</td>
<td>0.003</td>
<td>0.006</td>
<td>0.003</td>
<td>0.003</td>
<td>0.002</td>
<td>0.01</td>
<td>0.001</td>
<td>0.001</td>
<td>0.03</td>
<td>0.004</td>
</tr>
<tr>
<td>9</td>
<td>0.0002</td>
<td>0.001</td>
<td>0.12</td>
<td>0.002</td>
<td>0</td>
<td>0.003</td>
<td>0.0002</td>
<td>0</td>
<td>0.001</td>
<td>0.004</td>
</tr>
<tr>
<td>10</td>
<td>0.0002</td>
<td>0.001</td>
<td>0.12</td>
<td>0.02</td>
<td>0</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0</td>
<td>0.0002</td>
<td>0.004</td>
</tr>
</tbody>
</table>

E[Accidents] 0.8566 0.8 2.6462 1.7104 0.534 0.7537 1.3181 0.84 1.5702 1.502