The Limits of Foresight in an Uncertain World

Matthew J. Squair

Jacobs
64 Allara Street, Level 1 Block 21
Canberra ACT 2600
Email: matthew.squair3@jacobs.com

Abstract

From it's origins system safety has had to deal with the challenges of complex, novel and disruptive technologies. This paper explores the inherent limitations of the classical formulation of risk and proposes a four-quadrant risk model to better manage the challenges of novel and disruptive technologies.

Keywords: Uncertainty, risk, disruption, epistemic, aleatory, ontological, complexity, safety, overconfidence, ignorance

1 Introduction

> "*It is impossible to win the great prizes of life without running risks.*
>
> Theodore Roosevelt

System safety as a discipline emerged in the mid part of the last century as a response to the increasing complexity of systems, and the novel technologies that were then being developed, and we find that the challenges that these early practitioners faced are indeed the same that we face today. That is, when dealing with complex and novel technological systems how can we assure ourselves that we have managed the risks effectively? One of the classical tools of the discipline has been the conceptualisation of risk as a way to determine the acceptability of technological systems. Nonetheless despite their centrality to the practice of the discipline; risk, probability and what 'exactly' we mean by these terms continue to be problematic, and their problematic nature goes mostly unremarked within the discipline of system safety.

This paper argues that the classical definition of risk imperfectly captures the full range of uncertainty that the discipline of system safety deals with, and that we are better served by a broader operationalised definition of risk across a spectrum of uncertainty. In this paper we firstly review the classical formulation of risk and identify problems in it's application to system safety. Various ad hoc measures traditionally used to work around these problems are identified, and a broader spectrum of uncertainty and risk is established. From this an operational definition of risk is developed and used to develop the four quadrant risk model.

2 The origin of risk and it's problems

> "*The Risk of losing any sum is the reverse of Expectation; and the true measure of it is, the product of the Sum adventured multiplied by the Probability of the Loss.*"
>
> Abraham de Moivre

The formal mathematical treatment of random systems is generally agreed to have started with an exchange of letters between the mathematicians Fermet and Pascal (1654) on how to fairly divide up the points from an interrupted game of chance. The solution they came up with was the idea of an ensemble that is the set of all the potential outcomes of the game with the winnings being allocated on the basis of the proportion of games the players won. From this work de Moivre (1710) further developed the concept of using an ensemble of *parallel worlds* to model possible future events, *set theory* to define the probabilities of outcomes, and the combination of probability P and consequences S to define the *expectation* of loss, or risk R as it's inverse.

$$R = S \times P \tag{1}$$

This means that as risk is a scalar quantity we can in principal sum the ensemble of n system risks as follows.

$$R_T = \sum_{i=1}^{n} S_i \times P_i \tag{2}$$

Implicit in the classical-mathematical definition of risk is that probability and consequences are objective, knowable, and quantifiable, at least in principle. Certainly this is true for the closed system of Fermet and Pascal's card game, but is it always true? To answer that question we first need to consider what risk 'is'.

2.1 What is risk, really?

Although classically risk is defined as a mathematical quantity it does not represent a discrete physical property, given the impossibility of directly measuring a future state. Numerically calculated risk is simply not the same as the calculated value of the yield strength of a bar of steel, although we may wish to treat it as such (Downer 2017). Rather it is the combination of two estimates, that of the magnitude of potential loss and it's probability. After Bridgman et al. (1927) these values are *operationally* defined, that is the processes used to arrive at the values defines them. In the

A preprint of this paper was presented at the Australian System Safety Conference (ASSC2024), Brisbane, Australia.

case of classical risk that process is Pascalian calculus.

The inherently operational nature of risk naturally results in differing definitions of risk. For example the frequentist (empirical) definition of probability and the subjectivist (Bayesian belief) definition can and do result in different formulations of risk, for example quantitative F/n curves versus subjective risk matrices. Even small changes, such as altering how we measure exposure, can significantly affect how we perceive risk. For example calculating accident rates per kilometre can result in quite a different rate to those calculated per trip, depending on which unit of exposure better correlates with accidents (Weir 1999). Under conditions of uncertainty risk proxies, that is non-classical formulations of risk, may also be used as more robust estimators of risk. For example, during early system design risk measures based on the ranked effectiveness of risk controls may be a more robust risk estimator as opposed to probabilistic risk, given the uncertainty of probability estimates and our unawareness of potential failure scenarios or causal factors (Leveson 2019).

It seems then that there is no singular definition of 'risk' itself but rather various models and measures that serves to operationalise our necessarily subjective and incomplete *perception* of risk in the world. As Feyerabend (2020) once remarked about scientific progress, for risk theory it seems that 'anything goes'. We now consider the question of whether classical risk as a measure actually correlates with the risks that we experience, and care about, in the real world.

2.2 Ergodicity and commensurability

Returning to classical risk for a moment, the way in which Fermet and Pascal (1654) dealt with the uncertainty of outcome in their card game was to imagine a set of parallel worlds that comprised all the possible end states for their game then calculate the ensemble statistic for each player in terms of the proportion of hands they would have won. This parallel worlds approach if used in the context of an ensemble of independent decisions, such as Fermet and Pascal's card game, works quite well. However, there's another way to view decisions on risk and that is as a series of events in time, and how the sequence of events might affect an individual decision maker as they occur *in this world only*. To capture the difference between these two viewpoints, and their differing statistics requires us to introduce the concept of ergodicity.

Classical risk in it's use of the ensemble statistic inherently assumes the property of ergodicity, i.e. that the ensemble statistic is the same as the time series statistic for some system. Unfortunately we live in an unforgivingly non-ergodic world in which the ensemble and series statistics do in fact statistically diverge. For example if we look at the expected survival rate for $n \to \infty$ individuals playing a game of Russian roulette, we find that the average (ensemble statistic) settles on a 5/6 survival rate while for a single individual playing it $n \to \infty$ the expected survival rate or time average (the time series statistic) settles on 0. The parallel world model and the resultant ensemble statistic are clearly misleading when we consider catastrophic risks to an individual exposed to the risk sequentially in the single world they inhabit. This example illustrates unambiguously how using the classical formulation of risk can be be dangerously deceptive because of the built in risk neutrality/ergodicity assumption. That is the assumption that high consequence low probability events are the same as low consequence high events. This assumption, as the Russian Roulette example illustrates, breaks down in the presence of catastrophic (or non-ergodic) risks.

We also find that ergodicity is unevenly distributed across likelihood, by their very nature high consequence risks are both rare and non-ergodic while high frequency risks are ergodic and invariably minor. Unsurprisingly humans exhibit consistent risk aversion to non-ergodic risks when compared to the classical risk neutral calculus of experts (Slovic and Weber 2002; Ellsberg 1961). Unfortunately owe sometimes have difficulty in perceiving the difference between these risks which can result in perverse effects. Such perverse outcomes are illustrated by the so called fence paradox in which risk controls that eliminate low consequence (ergodic) risks reduce a decision makers perception of risk, resulting in more risky behaviour and an increased exposure to non-ergodic risk (Cirillo 2018; Adams 2006).

This problem of ergodicity results in two consequences for managing risk, the first is that risk acceptance based on risk neutrality underestimate the risk aversion of those exposed to high consequence risks. This results in ad hoc fixes of risk acceptance thresholds for risks of catastrophic potential, see for example the US DoD's venerable system safety standard MIL-STD-882 and it's nominal risk matrix (USAF 2012). The second is that for high consequence risks acceptability does not actually turn on whether a risk is acceptable. Instead, after Borel (1950), we are arguing that an event is so unlikely it has negligible effect and may therefore be disregarded. Borel uses the example of a Parisian being killed in a car accident to illustrate his 'argument of negligible effect'. Using the then annual traffic death rate of 1 per million in Paris Borel argued that any event of probability of 10^{-6} can be considered negligible for any one individual and that 10^{-15} could be considered negligible at the global (world) level.

This is why industries such as nuclear and aviation express safety in terms of accident rates not risk and why when AREVA, the designer of the European Pressurised Reactor (EPR), estimates the likelihood of a reactor core damage event as being 6.1×10^{-7} per year, they are not arguing that some quantified level of residual risk is 'acceptable' rather they are arguing that the event is so unlikely that it can be ignored (Ramana 2011). Of course, to make such an argument one also needs to be very certain of the numbers, and we turn now to the question of uncertainty and the probability of extreme but rare events.

2.3 Uncertainty and extreme events

The calculation of risk in the classical sense relies on us knowing the set of risk scenarios and their probabilities with precision sufficient to be able to bound the worst case and then calculate a specific ensemble value. This is normally achieved by assuming some central tendency for severity and a thin tail distribution in which extreme events may be neglected due to their low probability. If we cannot put an upper limit of severity in such a fashion then we are dealing with a fat tail distribution where extreme events come to dominate the risk ensemble, and in the the worst case the total risk can become effectively infinite (Aven 2011). This bounding of risk is rarely straight forward for novel technologies, where many factors are ambiguous or open to interpretation. The decision as to what is the bounding worst case, sometimes termed truncating the tail, also becomes more uncertain as the number of extreme tail events in the sample decreases.

A study by Vesely and Rasmuson (1984) of uncertainty as it relates to error factors for probability estimates in nuclear safety assessments found a positive relationship between the rarity of the event and the error factor with common failures having an error range of 1.3-2 as compared to rare (extreme) accidents having an error range of 20-30 (see Table. 1). Given individuals generally desire the most catastrophic of events to happen the least, these same events also tend to be the ones which have the most uncertainty as to their likelihood. In such circumstances the most robust decision strategy may be to bound the possible severity through deterministic design measures, or to otherwise ensure the corrigibility (reversibility) of risk scenarios, should they eventuate. This is a lesson that the nuclear industry seems to have finally learned with their 4th generation reactor designs, which by design physically limit the potential severity of reactor accidents (Gauthe et al. 2017). By reducing our vulnerability to outlier event uncertainties, physically bounding the potential severity can have greater payoffs than classical calculations of quantitative risk would indicate.

Table 1: Declared uncertainties of probability estimates in nuclear plan safety assessments expressed as the ratio of the upper 95th percentile or confidence value to the median or 50th confidence value. Events are either *Initiating (I)* or *End states (E)* (Vesely and Rasmuson 1984).

Event	Type (I/E)	Per year /demand*	Error factor
Common failure	I	> 1	1.3-2
Higher value failure	I	10^{-3}	10
Single human error	I*	10^{-3}	10
Unavailability	E	10^{-4}-10^{-5}	4-10
Accident	E	10^{-3}-10^{-5}	3-6
External event	I	10^{-3}-10^{-6}	10-30
Unlikely major failure	I	10^{-7}	20-30
Complex human error	I*	10^{-7}	20-30
Extreme accidents	E	10^{-7}	20-30

2.4 Risk is what you make of it

In summary then risk serves to express our necessarily subjective perception of the uncertainty about a future undesired outcome, and the loss that might accrue from it's occurence. There is no singular definition of 'risk' itself but rather various models and measures that serves to operationalise our necessarily subjective *perception* of risk for the context of interest. Where it involves non-ergodic outcomes the classical mathematical formulation of risk break down and the principle of risk neutrality cannot automatically be assumed, i.e. that commensurability of high and low consequence risks exists. Similarly where there is uncertainty about probability, or whether all risk scenarios have been identified, classical risk is not a robust estimator of risk and other forms may need to be substituted.[1]

3 Uncertainty and risk

> *"We demand rigidly defined areas of doubt and uncertainty!"*
>
> Douglas Adams

As the future is unknowable thinking about risk inherently entails thinking about the various forms uncertainty about the future may take. In it's most general sense *uncertainty* can be defined as having limited knowledge about a subject. However, limited knowledge is not simply the absence of it, it can also result from the inadequacy of the information in our possession which may be inexact, unreliable or in conflict. This includes the opinions of experts who may themselves be in conflict. Uncertainty here is not just over the probability of some future event, but also the possibility for error, inaccuracy, and the use of discretion and judgement in the event's description and subsequent characterisation (Redmill 2002). At the extreme it captures the state of absolute ignorance. We now introduce the uncertainty continuum of Figure. 1 bounded at one end by complete knowledge and at the other end by complete ignorance.

3.1 Uncertainty as randomness

At one end of the spectrum of uncertainty there are those things that we know and are very confident that we know. For example there might be some parameter of a process that we are interested in. We then collect some evidence and on the basis of the evidence are confident about it's value. Note that our parameter may not be deterministic, it may also describe a stochastic process, for example frequency or probability of occurrence.

Classical analysis techniques such as Fault Tree Analysis (FTA) and Failure Modes and Effect Analysis (FMEA) are based upon our 'knowing' aleatory uncertainty, for example in the form of component failure rates, which allow us to assess the risks associated with systems in a quantitative fashion. However, such analyses also rely upon Hume's assumption of consistent condition, i.e. that the future will be consistent with the past (Hume 2016). For such an assumption to be valid for complex technological systems there also needs to be a set of engineering and operational practices that ensure that the new design and intended operating environment are similar to the past. We are in effect constraining such systems to be as closed and quasi-static as possible, rather than open and dynamic system with all the uncertainty and chance that that brings. Our confidence in the numbers is based not just on the statistics but also on a disciplined design process where only incremental improvement, usually based on the hard lessons of past accidents, are pursued and we can reasonably assume consistent condition (Downer 2017). The experience of Boeing, and the FAA, in addressing the unanticipated risks posed by lithium-ion battery technologies serves to illustrate what can happen when this disciplined and conservative process breaks down (Kolly, Panagiotou, and Czech 2013).

[1] This concept draws on the field of robust statistics where a robust statistical estimator is insensitive to outliers and small departures from underlying model assumptions (Hampel 1971).

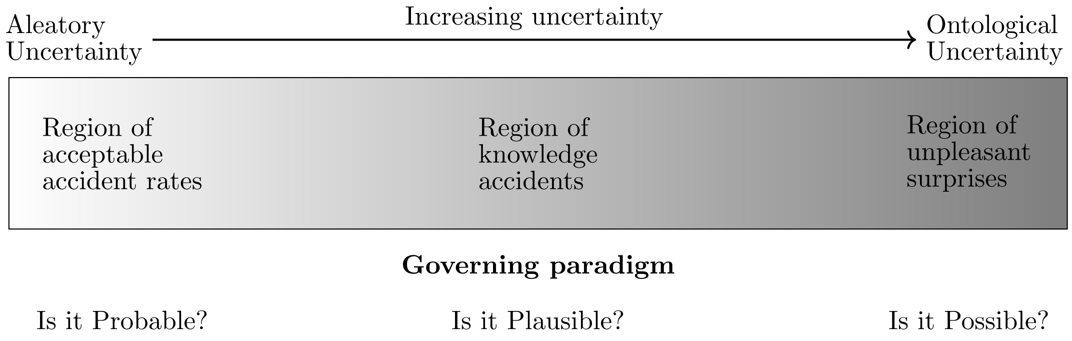


Figure 1: The continuum of uncertainty

From a risk perspective if there is *aleatory risk* then if a system is operated for long enough (or with enough systems) it will eventually incur a loss. For ergodic risk acceptance in such circumstances is about whether the loss rate is acceptable, for some defined duration of exposure, and what resources we are willing to expend in order to achieve an acceptable level of risk. For non-ergodic or catastrophic risks it is whether we can argue the negligible effect of the risk based upon probability. When we talk about such risks we use the language, and paradigm, of *probability*.

3.2 Uncertainty about what we know

Uncertainty about what we know, or *epistemic uncertainty*, represents our uncertainty about the state of knowledge. This uncertainty becomes increasingly significant as we rely more upon theory driven models to risk in the absence of empirical data. For example we may have a number of different potential plausible risk scenarios but be uncertainty about the underlying causal mechanisms. There may be uncertainty about the structure of causal models that we build or the model parameter values that we use. In the extreme we may be unable to express a preference for one alternative. Indeed Knight (1921) in his work Risk, Uncertainty, and Profit characterised this as a situation of ‘unmeasurable uncertainty’ when making decisions in the real world, in comparison to the quantitative certainties of classical risk theory.

In principle this uncertainty and it’s associated risk can be reduced through obtaining greater knowledge but in practice this does not address the series of epistemic hurdles. These range from how do we judge the potential inadequacy of the information that we possess, which may be considered to be inexact, unreliable or in conflict, through to our dependence upon theory to establish the correctness of evidence. Worse yet we can find that as we gain more information it exposes that our understanding to be more limited than we thought, or that the system is far more complex than we had theorised.

In risk analysis recognised sources of epistemic uncertainty include:

- Subjectivity and cognitive biases.
- Uncertainty over parameters or model structure.
- Imprecision/ambiguity in risk scenarios.
- Complexity (non-linearity/non-determinism).
- Language imprecision and ambiguity.
- Disagreement of information and expert opinion.
- The use of assumptions (Duhem-Quine).
- The practicality vs fidelity trade off.
- The ambiguity of ‘facts’ without theory.
- The failure to fully characterise losses.

For safety analyses this uncertainty is expressed over model parameters, such (aleatory) failure rates, and (epistemic) model structure. Here we are evaluating the effect of uncertainties upon the overall answer that we are seeking. For example in FTA we can apply Monte Carlo techniques to evaluate how parameter distributions affect a top level event’s probability distribution. Model structural uncertainty can be evaluated by robustness analysis that shows the extent to which outputs are driven by input rather than the internal structure, e.g internal model parameters and assumptions, and for any identified sensitivities identify which parts of the system model significantly effect the output and which do not (Young and Holsteen 2017). In a similar fashion to reliability analyses and their budgets, uncertainty analyses can be used to characterise uncertainty and budget allowable degrees of uncertainty to the elements of a system be analysed, for example through the allocation of integrity or assurance levels to functions and components. Alternatively we may elect to place claim limits on quantitative parameters such as system failure rates (Defence 1994).

Each of the sources of epistemic uncertainty requires it’s own set of independent measures to characterise, control and ideally reduce it during the risk analysis process. Imprecision may be addressed through careful definition of the terms of art, and risk scenarios while complexity can be addressed through deliberately pursuing simplicity for critical functions and processes. However, some types of uncertainty represent fundamental subjective bounds to risk analysis itself, for example the degree to which we can simplify a model, balance competing theories, or bound the analysis and therefor effort (while still deriving a usable answer) are essential decision making aspects of risk analysis, while also being inherently value laden and subjective (Morgan, Henrion, and Small 1998).

We may pursue system safety through robustness using satisficing strategies such as trading a small amount of performance for less vulnerability to violated assumptions or seeking system designs that perform adequately over a wide range of scenarios rather than optimally for a few. When we consider epistemic uncertainty we use the language and paradigm, of *plausibility.*

3.3 Uncertainty as ignorance

Ignorance represents a state of complete unknowing. Not only do we not know, but we don't even know what we don't know and we therefore have no direct perception of the true state of our uncertainty. Nassim Taleb proposed a category of risk to encompass this absolute, termed *ontological uncertainty* and associated with *black swan* events (Taleb 2005). Unfortunately when developing a new technology we are necessarily ignorant of many aspects of it, and some of these may hide as yet unidentified catastrophic risks. As a result we should not be surprised by disastrous surprises in new technologies, given the possibility for such events to be hiding within that uncertainty. The early history of high pressure steam boilers and of electricity are examples where technological ability out-paced an understanding of the dangers of the new technology. In both cases early attempts to limit the technology in order to control the risk proved unsuccessful, and it was eventually a combination of improved understanding of how failures and accidents could occur coupled with strict design and operation standards that resulted in improvements to safety (Leveson 1992). Even when we believe we understand the risks well we may still come to grief, for example the early focus of engineering for the 3rd generation digital fly-by-wire aircraft was upon the assurance of the dependability of the flight control systems, in reality the causes of major accidents in service were predominantly due to poor integration of crew and automation.

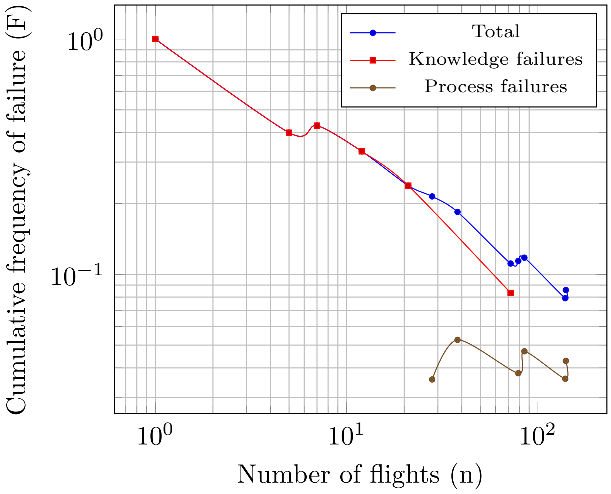


Figure 2: Knowledge and process failure rate for the RL-10 1962-2005 (Go 2008).

In another earlier example of the uncertainty inherent in new technologies Leonard Bostwick, an experienced NASA engineering manager involved in the H1 and F1 engine development for the Saturn booster program, noted that as a result of differing types of uncertainty in rocket engine development, these programs moved through four distinct phases each reflecting the dominance of a specific type of uncertainty (Bostwick 1968). Initial problems, and failures, occurred during early development because of the inability to totally extrapolate and build on existing knowledge, with such problems becoming evident only during full power trials, with usually catastrophic consequences. As the engines passed through subsequent integration and manufacturing and operational phases these in turn introduced risks peculiar to that phase, and indeed a subsequent study of NASA's LH2 RL-10 engine development by Go (2008) justifies his predictions. Figure. 2 derived from Go (2008) compares knowledge-based failure rates, due to incomplete understanding of the environment or the behaviour of the system, which decrease rapidly during initial operation against a relatively steady rate of process failures, resulting from procedural errors in manufacturing, verification and ground handling.

Conversely if a significant amount of time in operation is being spent on unplanned modifications and component changeouts to address anomalies and concerns found during actual system use, it's a good indicator that the system overall is also carrying a significant amount of developmental related uncertainty, and risk, into service. For example in the lead up to the 1986 Challenger disaster almost half the US Shuttle's turnaround time was devoted to addressing unexpected problems and anomalies (Pincus 1986). Unsurprisingly, given the (currently) novel nature of the technology, anomaly detection is an ongoing research area for ML both as a means to warn operators of potentially hazardous scenarios as well as to detect potential malign manipulation.

We can reasonably conclude that the early development of systems incorporating new technologies will face knowledge risks related to the incompleteness of 'what we know' when applied to beyond the state of art problems, with no magical road to success. This lesson should probably be borne in mind when we try to deploy Machine Learning (ML) while still arguing about foundational issues of ML safety, such as what actually constitutes 'interpretability'.

For systems that exhibit complex software behaviour it also appears their inherent complexity may also obscure ontological 'holes' resident in the system design. For example in a study of software maintenance Adam (1984) found a heavy tailed distribution for failures where approximately one third of faults only caused failures at the rate of once every 5000 years of execution, as Figure. 3 illustrates. In a subsequent study of software failure Bishop (1993) concluded that rather than a constant failure rate operational failures of software come in 'bursts' interspersed with long periods without failure, and that if software is operating successfully the probability of continued operation will be increased when there are only small changes in input conditions. Conversely failure probability would increase if there were major changes across the set of external conditions e.g. mode changes, hardware failure recovery or contingency operations. In a study of flight control software Hecht (2012) found that critical failures of complex software were due to software's inability to handle rare unspecified combinations of conditions, and that these failures were more likely to propagate to a major failure because they were unanticipated. Koopman reached a similar conclusion in regard to the challenge of *edge case* environmental scenarios to ML for autonomous vehicles, whose the heavy tail distribution of such rare events imposed an epistemic limit upon the efficacy of testing ML software for freedom from faults prior to deployment (Koopman and Wagner 2016). Thus an epistemic bound on our understanding seems to be an inevitable consequence of our

Table 2: Surprisal criteria of Parker and Risbey

Criteria	Results in	Example
Complexity	Difficulty in identifying knowledge gaps	The inscrutability of complex software
Novelty	Existing knowledge cannot be applied easily	Bostwick's first development stage Tacoma narrows span/width metric
Overconfidence	Unexpected conditions, events or anomalies	Boeing/FAA's certification strategy
Past surprises	Pre-existing unknown unknowns	Boeing battery reliability issues

building robust complex systems under time and resource constraints, with our resultant system being demonstrably robust, right up to the point where it runs into a rare and unanticipated event (Carlson and Doyle 2002).

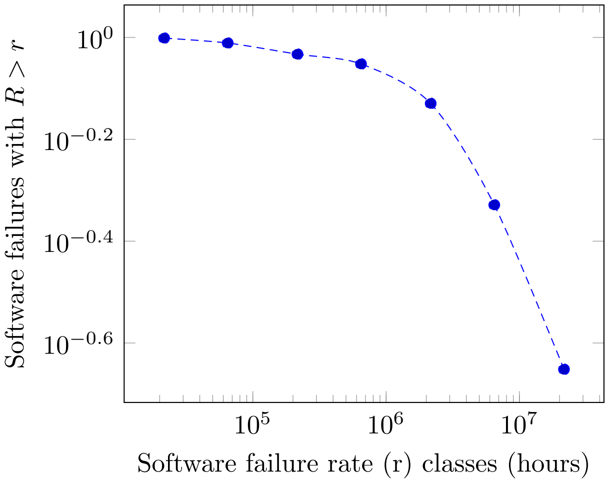


Figure 3: Cumulative plot of software failure rate classes. Note that the right hand side of the curve (very rare events) is sample censored by the total operational time (Adam 1984).

It appears inevitable then that where we have complex and critical behaviour persistent unidentified risks will exist waiting for the rare set of conditions to emerge and surprise us when we deploy these systems into behaviour rich environments. The QF72 Learmouth in-flight upset provides a canonical example of such an unpleasant surprise. The upset, a false triggering of stall protection laws and subsequent un-commanded pitch down of the aircraft, was caused by bursty ($n > 1$) noise that defeated the triple redundant median voting of the aircraft's Angle of Attack (AoA) sensor channel. That this failure occurred after hundreds of operating years of service for the aircraft type illustrates the 'robust but fragile' behaviour of our systems in the face of rare events for which a response has not been specified. In this case the presence of bursty (clustered) noise in the channel (ATSB 2013). The resultant extreme response was a result of the behaviour of median voting algorithms, which work adequately for single failures but can also generate extreme responses when presented with multiple failures (Bass, Latif-Shabgahi, and Bennett 1997).

At it's simplest when faced with deep uncertainty a precautionary approach to safety dictates that we take action with little evidence to guide us. This is captured in the concept of *precaution*. For example had Boeing designed it's battery box such that a single cell over temperature did not result in a cascading failure of adjacent cells, the failure to identify hazardous scenarios would have resulted in a reliability issue rather than a fire with potentially catastrophic consequences. The precautionary approach also extends into the operation or deployment of a system where the occurrence of anomalies should be treated as an indication that there is a gap in our knowledge and trigger investigation until we are satisfied we understand it's causes, and consequences. Such an approach differs from traditional engineering methods of understanding systems as a whole, instead adopting what we might term the ecologist's approach to understanding large scale complexity, that is seeking to understand a specific small part of the whole in order to infer small scale but still useful generalisations (Arbesman 2017). For highly complex systems a 'sense-making' approach of probe-sense-response may be better suited to managing risks than trying to understand the whole. When we consider deep or ontological uncertainty we use the language and paradigm, of *possibility*.

In evaluating the risk associated with deep uncertainty classical risk formulations are of little assistance, instead we must develop measures that provide us with criteria, such as that of Tacoma, that provide a measure of ontological uncertainty and the possibility of unpleasant surprises. Rather than attempting to predict new risks from a position of deep uncertainty we can adopt a *horizon scanning* approach of attempting to detect early signs of potentially unidentified risks through a systematic examination of new technology, the uncertainties involved and their effects.[2] Horizon scanning requires an explicit focus on the early detection of 'weak signals' as indicators of potential change or risk, and is normally implemented by reviewing near misses, anomalies, end-user feedback and research into how others had implemented like technologies. As an example, Boeing 787 aircraft battery fires were preceded by a series of unexpected service failures and anomalies, while both Boeing and the FAA appeared to have ignored the considerable experience gained by the vehicle industry in dealing with lithium battery fire risk. Likewise, the lessons learned in the development of autonomous vehicles have broader application than just vehicles.

The Tacoma Narrows Bridge disaster provides an example of how we can, in principle, establish practical measures for the degree to which our belief in a design hypothesis is justified. To reduce the cost of constructing the Tacoma bridge the designers applied a novel structural design hypothesis that took account of the ability of the suspension cables to transfer horizontal wind loads to the pylons. This allowed the use of shallower and cheaper I-beams to stiffen the roadway rather than traditional, more expensive, trusses (Petroski 1992). However, the long slender bridge was also much more flexible allowing the wind to ex-

[2] Horizon scanning can also act as a defence against near miss bias amongst decision makers and provide a window into otherwise opaque system behaviour.

cite a catastrophic torsional flutter mode causing it's destruction. Formulating the risk of design related failure in probabilistic terms provides no real insight into risk as we have nothing to substantiate a numerical probability. However, we can as Table. 3 illustrates, formulate a simple comparison of the span to width ratio of Tacoma with other suspension bridges. Such measures can inform us as to how far outside the experience base we are were working, in the case of Tacoma quite a lot.

Table 3: A comparison of the Tacoma Narrows bridge width to span ratio with other bridges.

Year built	Bridge name	Width/Span
1849	Wheeling	1:16
1929	Ambassador	1:31
1939	Whitestone	1:31
1931	San Fransisco	1:35
1931	Golden gate	1:47
1940	Tacoma Narrows	1:72

Although it is impossible a-priori to come up with singular measure of uncertainty, given the uniqueness of each system, it is possible to at least categorise the various ways in which an unjustified belief supports a design, as a useful metric of how much risk of unpleasant surprises we may be exposed to. To that end Risbey and Parker et al. developed a set of characterisation criteria, summarised below in Table. 2, to provide insight into the degree to which we might expect surprises. As Parker and Risbey (2015) point out any assessment of uncertainty should take account of all significant sources, and should consider all available (relevant) information when doing so. In the case of deep uncertainty that includes consider the possibility of encountering surprises. The authors also cautioned against adopting a once size fits all approach to uncertainty assessments and the inappropriateness of using probabilistic measures for deeper uncertainty as this expresses an unwarranted degree of precision.

3.4 Uncertainty and corrigibility

As Collingridge (1982) points out, novel technological systems present us with a double bind decision dilemma in dealing with novel technologies . Initially we are presented with a problem of deep uncertainty because 'real' issues and their risk cannot be easily identified or quantified until a new technology is deployed.[3] But once a new technology (or system) is deployed we then face the problem of *path dependence* where making significant changes is very difficult. Rittel and Webber point out that such irreversibility becomes a wicked planning problem, where attempts to reverse the decision will carry with them their own set of intractable problems, and be subject to the same effects (Rittel and Webber 1973). Resolving this paradox is one of the central challenges of managing the safety of emerging technologies and novel systems.

The approach recommended by Collingridge was to pursue the corrigibility of decisions to commit to a particular technology, by ensuring that such decisions are reversible. It is generally better to take a series of incremental steps not committing to the next step until the risk is better understood, or as Deng Xiaoping once put it, "sometimes you have to cross the rivers by feeling the stones" (Browning and Boudès 2005). Returning to the concept of ergodic and non-ergodic risks again, we can characterise corrigibility strategies as moving risk from non-ergodic to ergodic through reducing the irreversibility of consequences.

System architectures in combination with a compatible development lifecycle can be used to achieve corrigibility when deploying a new technology. For example low investment precursor systems can be used to explore technology and mature them before fully committing to them. Architectural strategies such as the simplex architecture proposed by Seto, Danbing et al. (1998) may be used to provide both an ability to add new system functionality whilst simultaneously providing high assurance of being able to revert to safe behaviour if an unsafe state is detected.

At the functional design level where a system intentionally incorporates irreversible behaviours the associated non-ergodic risk can be reduced by providing many paths to safe states and fewer paths to hazardous states making it easier to transition to a safe state and harder to transition to a higher and more hazardous state.

In combination with the ability to interrupt a system's behaviour corrigibility (reversibility) forms the meta-property of system controllability. Controllability has long been recognised as an essential aspect of system behaviour in classical control systems theory. Traditional control theory also teaches us that controllability is particularly difficult when we deal with non-linear systems, because reversibility cannot be guaranteed thereby providing a strong system theoretic justification for avoiding non-linear system behaviour in safety critical systems.

3.5 Uncertainty margins

Returning to classical risk assessments which attempt to quantify a quantitative frequency of loss or other risk metric the presence of unknown and therefore unquantified risks is an obvious confounding factor. To address this we can apply a traditional engineering technique and establish an uncertainty margin U to allow for unidentified risks, that is then deducted from the quantitative risk criteria R_c, such as a maximum allowable accident frequency, to establish the quantitative risk requirement R_r giving Equation. 3 (Dezfuli et al. 2015). The size of this margin can be set based on the degree to which novel technologies are being used, system complexity might obscure causal factors or we believe the operational environment may contain surprises.

$$R_r \leq R_c - U \tag{3}$$

By incorporating uncertainty explicitly into our risk criteria we can quantify how we can mitigate total risk by reducing uncertainty (and possible unknown risks) or by reducing identified risk through explicit risk controls. In principle this allows us to quantitatively represent the:

- Trade-offs between known and unknown risks.
- Reduction of unknown risk over time.
- Introduction of unknown risk via changes.[4]

[3] Collingridge was gloomy about our ability to work on the 'prediction' side of the dilemma even using a Bayesian approach, due to the high level of uncertainty that in his view invalidated Bayesian risk assessments.

[4] This includes changes made to mitigate risks.

- Correlation of unknown to known risks.
- Growth of safety over time (ideally).

By including a parameterisation of uncertainty into risk we are of course accepting that the resultant R_r contains a large degree of uncertainty and pursuing a satisficing risk strategy, i.e. we are looking for a satisfactory (but not perfect) answer in the context of a more complete (but less precise) appreciation of risk (Simon 1956).

4 Risk, uncertainty and decision making

> "*...there are things that we know we know, there are things that we know we don't know and then there are things we don't know we don't know.*"
>
> Donald Rumsfeld

In the preceding sections of this paper we introduced the concept of a spectrum of uncertainty and for the purposes of discussion divided it into a number of domains. That these types and degrees of uncertainty exist is significant because they affect the decision making strategy we may adopt (Lempert and Collins 2007). For example if risk can be precisely defined and quantified then classical utility based decision making may be successfully applied, if significant epistemic uncertainty exists then applying robust (non-predictive) decision strategies would be more appropriate, while if we believe there are deep ontological uncertainties in our understanding then a precautionary strategy would be most appropriate.

4.1 The four quadrant model

Translating Rumsfeld's aphorism into four quadrants provides a useful framework onto which we can map the various forms of uncertainty and their associated paradigms. In this model each quadrant represents a different form of uncertainty with it's own paradigm, dimensions of uncertainty and associated decision making strategies that should be applied. The quadrants are summarised as follows:

- Q1. Unknown knowns. A mirror to epistemic uncertainty. This representing the circumstances where information may be known but is not available to decision makers. 'Could we have known' is the paradigm.
- Q2. Known knowns. Aleatory uncertainty and probabilistic risks. The domain of utility based decision making, design standards, quantitative safety and reliability engineering. 'Is it probable' is the uncertainty paradigm.
- Q3. Known unknowns. Epistemic uncertainty, there are unknowns but we have identified them in some fashion. The domain of robustness against anticipated failure and off nominal conditions. Potentially reducible through data gathering and research. 'Is it plausible' is the uncertainty paradigm.
- Q4. Unknown unknowns. Deep ontological uncertainty reducible through fundamental research. The domain of precaution, resilience and horizon scanning. 'Might it be possible' is the uncertainty paradigm.

4.2 Risk as lost knowledge

While we have previously discussed uncertainty and risks that map to quadrants two to four quadrant one represents those circumstances in which the information needed to make a correct decision or to effectively manage a risk is present, and in principal available, but the decision maker is not aware of it or cannot access it. The Boeing battery fire case study illustrates risks in this quadrant, where the FAA and Boeing's certification strategy would have benefited from a more rigorous investigation of the lessons learned in non-aviation domains such as electric vehicles, such as those promulgated by the US National Fire Protection Research Foundation (NFPRF) (Mikolajczak et al. 2012).[5] This quadrant also captures circumstances where decision makers may be acting locally in a rational fashion, based upon their local knowledge, but at a global level appear to be acting irrationally relative to organisational knowledge that they are not aware of (Vlaev 2018).

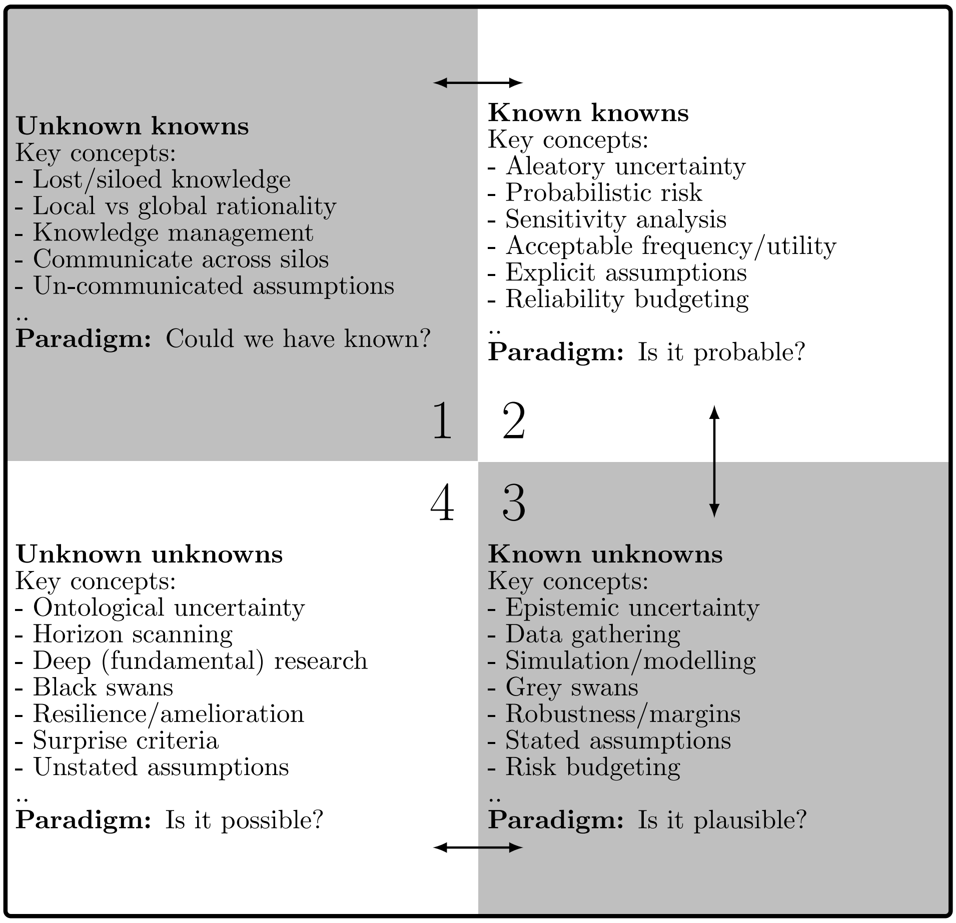
4.3 Risk and anti-fragility

As the Figure. 4 illustrates risk can also transform from one form to another through our actions to reduce and transform the attendant uncertainties. In aviation for example innovation is still achieved through initially deploying new technologies in the military domain, with it's much greater willingness to trade risk for capability, with uptake by the commercial aviation occurring once risk associated with the uncertainties of the technology being reduced to an acceptable level. Technologies that are viable, composites for example, then find their way into civil aviation and are slowly integrated into the stable design baseline. This approach allows the overall aviation system to achieve what Taleb (2012) terms *anti-fragility* that is an ability to learn from and grow in the face of accidents and failures.

This process can indeed often throw up solutions to hazards that the commercial aviation industry is unaware, for example the military had grappled with the problem of unreliable air data, and developing robust estimator alternatives to analog sensing during air combat manoeuvring has been a busy research area for many years in contrast to the civil aviation community (Mitchell 2004). Had such technologies been adopted by civil aviation accidents involving unreliable air data, such as QF72 and Air France AF447 may not have occurred.

At this point we should also note that risk flows can be bi-directional. For example we can certainly reduce the (aleatory) risk of random failure of a flight control system through the use of active redundancy, while unintentionally introducing the risk of undetected design flaws in the complexities of the resultant redundancy scheme, whose likelihood cannot be expressed as a frequency but is instead conditional upon the occurrence of the rare combination of events that would trigger them, as occurred for QF72 (ATSB 2013). We may also elect to explicitly trade off risk in the other direction. For example although solid state switches are more reliable, in safety critical functions electro-mechanical switches are traditionally preferred to solid state switches. On the face of it this seems an irrational decision,

[5] After Rae et al. (2014) we may also view this as a specific example of *probative blindness* in that efforts were narrowly focused upon battery design compliance versus a broader understanding of potential hazards.



Unknown knowns
Key concepts:
- Lost/siloed knowledge
- Local vs global rationality
- Knowledge management
- Communicate across silos
- Un-communicated assumptions
..
Paradigm: Could we have known?

1

Known knowns
Key concepts:
- Aleatory uncertainty
- Probabilistic risk
- Sensitivity analysis
- Acceptable frequency/utility
- Explicit assumptions
- Reliability budgeting
..
Paradigm: Is it probable?

2

4

Unknown unknowns
Key concepts:
- Ontological uncertainty
- Horizon scanning
- Deep (fundamental) research
- Black swans
- Resilience/amelioration
- Surprise criteria
- Unstated assumptions
..
Paradigm: Is it possible?

3

Known unknowns
Key concepts:
- Epistemic uncertainty
- Data gathering
- Simulation/modelling
- Grey swans
- Robustness/margins
- Stated assumptions
- Risk budgeting
..
Paradigm: Is it plausible?

Figure 4: The four quadrant risk management framework

but if we look at the decision from the perspective of uncertainty we find that the failure modes of solid state switches are more uncertain leading to uncertainty as to their modality and likelihood. Hence the rational decision to trade off increased component failure rates or aleatory uncertainty, against reduced epistemic uncertainty regarding failure modes.

The model also allows us to identify circumstances in which we are dealing with more than one form of uncertainty, for example in attempting to certify the Boeing lithium battery box design the FAA was in reality dealing with with two types of uncertainty, one explicit and one implicit. The first explicit uncertainty was whether the battery system would fail, expressed in probabilistic terms as no more that 10^{-7} failures per flight hour. The second implicit and deeper uncertainty related to the adopted certification process and it's ability to provide appropriate confidence that Boeing had *actually* achieved the required failure rate. In the absence of such scrutiny Boeing and the FAA fell victim to a form of experimenter's regress, where the the adequacy of the agreed certification basis was implicitly assumed, and testing subsequently confirmed that the battery design met the conditions, although the results did not in fact validate the initial assumptions. Engineering facts rather than being objective truths are perhaps better understood as *knowledge claims* that may be tested in the laboratory but in the end are validated by their practical use. Under conditions of uncertainty rather than 'rushing to certify' a more robust approach is to apply strategies appropriate to the level of uncertainty to characterise and reduce it where possible, instead of pretending greater certainty than is warranted.

5 Conclusions

The difference between the various types of uncertainty in developing novel technological systems goes some of the way towards explaining both our initial successes in system safety and the challenges that currently face us. Early practitioners of system safety focused upon the aleatory risks posed by the random failure or wear-out of critical system components and developed techniques such as FTA and FMECA to analyse the systems for these component behaviour derived safety properties. And indeed through the improvement of reliability, and the use of redundancy and design margins to handle system and environmental perturbations significant gains in safety have been made. But as the safety of systems improved through the reduction in aleatory risk the residual risks borne of deeper uncertainties became a proportionally greater cause of system accidents. At the same time systems have grown in both complexity and size thereby making it much more difficult to identify such risks and compounding these residual uncertainties.

In such circumstances it is useful, indeed necessary, to reformulate the classically derived formulation of risk within a broader understanding of uncertainty along with the appropriate analysis and management strategies. Refactoring the spectrum of uncertainty into a four quadrant model provides a useful tool with which to conceptualise our understanding of the full spectrum of risks that we face.

References

Adam, E.N. (1984). "Optimizing preventive service of software products". In: *IBM Journal Research and Development*. Vol. 28. 1. IBM, pp. 2–14.

Adams, John (2006). "The failure of seat belts legislation". In: *Clumsy Solutions for a Complex World*. Springer, pp. 132–154.

Arbesman, Samuel (2017). *Overcomplicated: Technology at the limits of comprehension*. Penguin.

ATSB (2013). *In-flight upset - Airbus A330-303, VH-QPA, 154 km west of Learmonth, WA, 7 October 2008*. Accident Investigation Report AO-2008-070. PO Box 967, Civic Square ACT 2608 Australia: Australian Transportation Safety Bureau.

Aven, Terje (2011). *Misconceptions of risk*. John Wiley & Sons.

Bass, Julian M, G Latif-Shabgahi, and Stuart Bennett (1997). "Experimental comparison of voting algorithms in cases of disagreement". In: *EUROMICRO 97. Proceedings of the 23rd EUROMICRO Conference: New Frontiers of Information Technology (Cat. No. 97TB100167)*. IEEE, pp. 516–523.

Bishop, Peter G (1993). "The variation of software survival time for different operational input profiles (or why you can wait a long time for a big bug to fail)". In: *FTCS-23 The Twenty-Third International Symposium on Fault-Tolerant Computing*. IEEE, pp. 98–107.

Borel, Emil (1950). "Les probabilités et la vie". In.

Bostwick, L (June 1968). "Development of LOX/RP-1 engines for Saturn/Apollo launch vehicles". In: *4th Propulsion Joint Specialist Conference*. Reston, Virigina: American Institute of Aeronautics and Astronautics.

Bridgman, Percy Williams et al. (1927). *The logic of modern physics*. Vol. 3. Macmillan New York.

Browning, Larry and Thierry Boudès (2005). "The use of narrative to understand and respond to complexity: A comparative analysis of the Cynefin and Weickian models". In: *E: CO* 7.3-4, pp. 32–39.

Carlson, Jean M and John Doyle (2002). "Complexity and robustness". In: *Proceedings of the national academy of sciences* 99.suppl 1, pp. 2538–2545.

Cirillo, Pasquale (2018). *Of risk, fences and unavoidable falls*. URL: http://www.pasqualecirillo.eu.

Collingridge, David (1982). *The Social Control of Technology*. St Martin's Press.

de Moivre, A. (1710). "'De Mensura Sortis' or 'On the Measurement of Chance'". In: *Philosophical Transactions (1683-1775)* 27.

Defence, UK Ministry of (1994). *Safety Management Requirements for Defence Systems*. Standard DEF STAN 00-56/1 Issue 2. UK MoD.

Dezfuli, Homayoon et al. (2015). *NASA system safety handbook. Volume 2: System safety concepts, guidelines, and implementation examples*. Tech. rep.

Downer, John (2017). "The aviation paradox: Why we can 'know' jetliners but not reactors". In: *Minerva* 55.2, pp. 229–248.

Ellsberg, Daniel (1961). "Risk, ambiguity, and the Savage axioms". In: *The quarterly journal of economics*, pp. 643–669.

Fermet, P. and B. Pascal (1654). *The Problem of the Points (Letters)*. Correspondance.

Feyerabend, Paul (2020). *Against method: Outline of an anarchistic theory of knowledge*. Verso Books.

Gauthe, P et al. (2017). "Considerations on GEN IV safety goals and how to implement them in future Sodium-cooled Fast Reactors". In: *International Conference on Fast Reactors and Related Fuel Cycles Next Generation Nuclear Systems for Sustainable Development (FR17)*. International Conference on Fast Reactors and Related Fuel Cycles Next ...

Go, Susie (Jan. 2008). "A Historical Survey with Success and Maturity Estimates of Launch Systems with RL10 Upper Stage Engines". In: *2008 Annual Reliability and Maintainability Symposium*. NASA Ames Research Center. IEEE, pp. 491–495.

Hampel, Frank R (1971). "A general qualitative definition of robustness". In: *The annals of mathematical statistics* 42.6, pp. 1887–1896.

Hecht, Herbert (2012). *Flight-Critical Systems Design Assurance*. Technical Report DOT/FAA/AR-11/28. Northwest Mountain Region – Transport Airplane Directorate, 1601 Lind Avenue, SW Renton, WA 98057: U.S. Department of Transportation, Federal Aviation Administration.

Hume, David (2016). "An enquiry concerning human understanding". In: *Seven masterpieces of philosophy*. Routledge, pp. 183–276.

Knight, Frank H (1921). *Risk, uncertainty and profit*.

Kolly, Joseph M, Joseph Panagiotou, and BA Czech (2013). "The investigation of a lithium-ion battery fire onboard a Boeing 787 by the US National Transportation safety board". In: *Safety Research Corporation of America: Dothan, AL, USA*, pp. 1–18.

Koopman, Philip and Michael Wagner (2016). "Challenges in autonomous vehicle testing and validation". In: *SAE International Journal of Transportation Safety* 4.1, pp. 15–24.

Lempert, Robert J and Myles T Collins (2007). "Managing the risk of uncertain threshold responses: comparison of robust, optimum, and precautionary approaches". In: *Risk Analysis: An International Journal* 27.4, pp. 1009–1026.

Leveson, Nancy (2019). "Improving the standard risk matrix: Part 1". In: *Cambridge, Mass.: Department of Aeronautics and Astronautics, MIT.*

Leveson, Nancy G (1992). "High-pressure steam engines and computer software". In: *Proceedings of the 14th international conference on Software engineering*, pp. 2–14.

Mikolajczak, Celina et al. (2012). *Lithium-ion batteries hazard and use assessment*. Springer Science & Business Media.

Mitchell, Eric John (2004). "F/A-18A-D Flight Control Computer OFP Versions 10.6. 1 and 10.7 Developmental Flight Testing: Out-of-Controlled Flight Test Program Yields Reduced Falling Leaf Departure Susceptibility and Enhanced Aircraft Maneuverability". Masters thesis. University of Tennessee - Graduate School.

Morgan, M G, M Henrion, and M Small (1998). *Uncertainty*. A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge: Cambridge University Press.

Parker, Wendy S and James S Risbey (2015). "False precision, suprise and improved uncertainty assessment". In: *Phil. Trans. R. Soc. A* 373, pp. 1–13.

Petroski, H. (1992). *To Engineer is Human, The Role of Failure in Successful Design*. First Vintage Edition. New York: Random House.

Pincus, W. (1986). *NASA's Push to Put Citizen in Space Overtook Fully 'Operational' Shuttle*. URL: https://www.washingtonpost.com/archive/politics/1986/03/05/nasas-push-to-put-citizen-in-space-overtook-fully-operational-shuttle/29fe2714-39b7-40dd-b15e-073441de636e/.

Rae, AJ et al. (2014). "Probative blindness: how safety activity can fail to update beliefs about safety". In.

Ramana, M V (Apr. 2011). "Beyond our imagination: Fukushima and the problem of assessing risk". In.

Redmill, Felix (2002). "Exploring subjectivity in hazard analysis". In: *Engineering Management Journal* 12.3, pp. 139–144.

Rittel, Horst WJ and Melvin M Webber (1973). "Dilemmas in a general theory of planning". In: *Policy sciences* 4.2, pp. 155–169.

Seto, Danbing et al. (1998). "The Simplex architecture for safe online control system upgrades". In: *Proceedings of the 1998 American Control Conference. ACC (IEEE Cat. No. 98CH36207).* Vol. 6. IEEE, pp. 3504–3508.

Simon, Herbert A (1956). "Rational choice and the structure of the environment." In: *Psychological review* 63.2, p. 129.

Slovic, P and E U Weber (2002). "Perception of Risk Posed by Extreme Events". In: *Risk Management strategies in an Uncertain World.* New York, pp. 1–21.

Taleb, Nassim (2005). "The black swan: Why don't we learn that we don't learn". In: *NY: Random House.*

Taleb, Nassim Nicholas (2012). *Antifragile: Things that gain from disorder.* Vol. 3. Random House Incorporated.

USAF (2012). *Department of Defence Standard Practice, System Safety.* Standard MIL-STD-882E. US Department of Defence.

Vesely, W E and D M Rasmuson (Dec. 1984). "Uncertainties in Nuclear Probabilistic Risk Analyses". In: *Risk Analysis* 4.4, pp. 313–322.

Vlaev, Ivo (2018). "Local choices: rationality and the contextuality of decision-making". In: *Brain sciences* 8.1, p. 8.

Weir, Andrew (1999). *The Tombstone Imperative: The Truth About Air Safety.* London: Simon and Schuster.

Young, Cristobal and Katherine Holsteen (2017). "Model uncertainty and robustness: A computational framework for multimodel analysis". In: *Sociological Methods and Research* 46.1, pp. 3–40.