

# Why Five 9s is not Five Stars – The Need for Out-of-Distribution Robustness Testing in AI Functional Safety

Erik Reynolds, Founder, Reynolds & Moore, Arya Gautam, AI Safety Engineer, Reynolds & Moore, Ferdaousse Ziari, Engineering Manager, Reynolds & Moore, and Ashley Weis, Senior Technical Writer, Reynolds & Moore

## Abstract

Quantifying the reliability of safety-relevant systems that incorporate Artificial Intelligence (AI) exposes common misapplications of reliability theory to systems that lack explainability. Traditional reliability engineering methods developed for simple or complicated hardware components depend on epistemic and statistical assumptions that do not translate to complex systems. Upon examination, these assumptions do not hold for AI. This document presents a framework for quantifying the reliability of AI-enabled safety-relevant systems.

*Index Terms*– Functional Safety, Artificial Intelligence, Safety Integrity Level (SIL), Diagnostic Coverage (DC), Probability of Dangerous Failure per Hour (PFH), Average Probability of Dangerous Failure on Demand (PFD<sub>avg</sub>), Safe Failure Fraction (SFF), IEC 61508, IEC 22440, PAS 8800.

[15], [16], [17], [18], [19], [20], [21]

## I. PURPOSE

This document provides guidance for technical personnel who design and integrate artificial intelligence (AI) elements within the framework of functional safety for electrical, electronic, and programmable electronic (E/E/PE) systems. Prevailing functional safety standards, including the International Electrotechnical Commission (IEC) 61508 and IEC 62061, have not historically addressed the quantification of AI reliability. This document differentiates traditional reliability engineering methods, which apply to non-complex components, and introduces a framework for complex systems reliability engineering that characterizes the contribution of AI elements to overall system reliability.

## II. INTRODUCTION

Safety functions that utilize electrical and electronic elements across diverse applications follow standards such as IEC 61508 and IEC 62061. These standards define lifecycle requirements, including planning steps for projects that apply functional safety, and provide guidance on allowable element-level failure rates for portions of safety function. New standards are in development to incorporate requirements for implementing artificial intelligence elements within safety-relevant systems.

Safety-relevant systems that incorporate AI models have the potential to enhance performance relative to traditional logic-based safety elements. Consistent with ISO/IEC Guide 51 terminology, these models support rapid detection of persons or

objects whose presence may lead to hazardous events based on leading indicators, detection of deviations between inputs and outputs, and advanced classification. In defined applications, these capabilities expand the functionality and usability of safety-relevant systems beyond traditional logic-based safety functions.

A primary challenge in applying artificial intelligence elements in safety-related systems is the opacity of these models. Standards such as IEC 61508 require comprehensive documentation and clear human understanding of system behavior, safety performance metrics, and failure rates. Because various models do not provide sufficient explainability, the use of AI in safety-related systems has often been restricted or prohibited.

Applying traditional reliability engineering frameworks wholesale to AI-enabled systems is inappropriate. Those frameworks rely on simplifying assumptions that enable practical probabilistic quantification, including independent and identically distributed failure data, constant failure rates, stationarity of element and system configurations, limited complexity of the operational environment, and explainable failure modes with transparent effects. These assumptions seldom hold for AI elements.

The remainder of this document proceeds as follows. Section III describes traditional reliability engineering assumptions. Section IV identifies challenges that artificial intelligence introduces for those assumptions. Section V presents a framework for quantifying the reliability of artificial intelligence. Section VI explains why wholesale application of traditional reliability engineering assumptions is unsafe for safety systems that incorporate AI.

## III. TRADITIONAL RELIABILITY ENGINEERING

In reliability engineering, Weibull analysis, also known as life data analysis, is a primary method for estimating failure rates. The process collects observations from a sample population of hardware element, subsystem, or system operating in a controlled or otherwise characterized environment. Recorded failures are then categorized by failure mode.

[Table 1] — Example of Failure Mode Distribution (FMD-91)

Device Type	Failure Mode	Failure Mode Probability
Electric Motor, AC	Winding Failure	0.31
	Bearing Failure	0.28
	Fails to Run, After Start	0.23
	Fails to Start	0.18

A plotting or regression procedure estimates the Weibull distribution parameters that best fit the observed data. Confidence bounds for these estimates are then computed at a specified confidence level. Goodness of fit is evaluated using

established methods, such as the chi-square test or the Kolmogorov-Smirnov test. The resulting parameter estimates define a Weibull model for the data, from which the failure function, reliability function, and hazard function are derived, enabling calculation of failure rates.

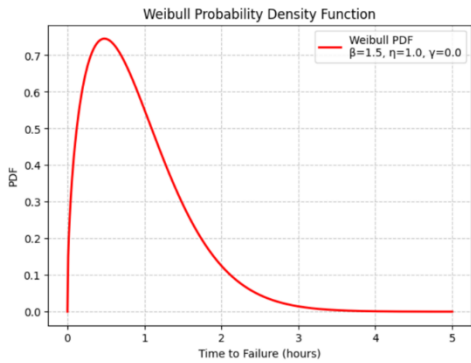
$$F(t) = 1 - \exp \left[ - \left( \frac{t - \gamma}{\eta} \right)^\beta \right] \quad [F1]$$

$$R(t) = \exp \left[ - \left( \frac{t - \gamma}{\eta} \right)^\beta \right] \quad [F2]$$

$$H(t) = \frac{F(t)}{R(t)} \quad [F3]$$

- $F(t)$  = Failure Distribution Function.
- $R(t)$  = Reliability Function.
- $H(t)$  = Hazard Function.
- $\beta$  = Shape Parameter *or* Weibull Slope.
- $\eta$  = Scale Parameter *or* Characteristic Life.
- $\gamma$  = Location Parameter *or* Expected Minimum Life.
- $t$  = Time.

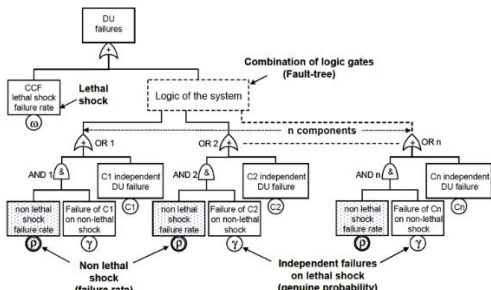
[Figure 1] – Hardware Component Weibull Plot



[24]

Failure mode and failure rate data are integrated to characterize overall system reliability. Common methods include Reliability Block Diagrams, Failure Modes and Effects Analysis (FMEA), Petri Nets, Markov Analysis, and Fault Tree Analysis (FTA).

[Figure 2] – Hardware System Fault Tree Analysis



[23]

Weibull analysis and related methods rely on epistemic and statistical assumptions. These assumptions make reliability engineering tractable for simple or complicated systems. The following section demonstrates that they do not hold for complex systems, including AI-enabled safety-related systems.

[Table 2] – Reliability Engineering Epistemic Assumptions

Assumption	Description
1	Failure modes can be known as <i>a priori</i> .
2	Failure mechanisms are transparent.
3	Failure behavior is explainable.
4	Representative populations are observable.

[Table 3] – Reliability Engineering Statistical Assumptions

Assumption	Description
1	Independent and identically distributed data.
2	Negligible impact of outliers on the tails.
3	Negligible change in variation over time.
4	Negligible complex interaction among variables.

#### IV. COMPLEX SYSTEMS RELIABILITY

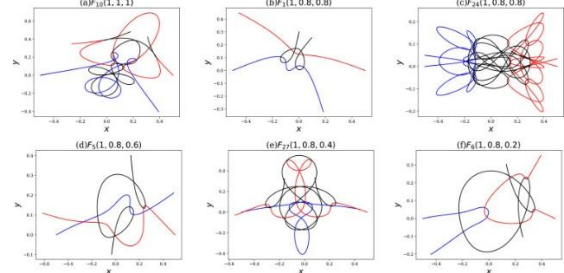
Complex systems are co-evolving, multi-layer networks that exhibit emergent behavior. They differ from complicated systems because this emergent behavior cannot be directly attributed to specific components. This behavior arises from interactions among components, which reduces explainability and limits transparency. This section examines how these properties of complex systems affect the epistemic and statistical assumptions that underline traditional reliability engineering.

##### 1) Epistemic Assumptions

###### a) Failure modes can be known as *a priori*.

Although complex systems comply with physical laws, small uncertainties in system state make emergent behavior indeterminable from analysis of the constituent parts. The classical three-body problem illustrates these limitations: Newtonian mechanics provides closed-form solutions for two interacting bodies, but no general closed-form solution exists for three bodies. Accordingly, the complete set of failure modes for a complex system cannot be known as *a priori* and is revealed only through observation of the integrated system across operating conditions and time.

[Figure 3] – Three Body Problem



[8]

*b) Failure mechanisms are transparent.*

The same sensitivity to small uncertainties in system state that prevents a priori enumeration of failure modes also limits a posteriori determination of failure mechanisms. Input-output relationships and pass-fail states are mappable, and intermediate model states may be examined through explainable artificial intelligence. However, these analyses do not consistently yield a complete or transparent causal account of failure mechanism; the mechanism remains only partially characterized.

*c) Failure behavior is explainable.*

For complex systems, this assumption does not hold. The unknowability of both failure modes and failure mechanisms makes failure behavior only partially explainable. Input-output relationships and pass-fail states are mappable, and intermediate states may sometimes be examined, but a definitive logical or chronological chain of causation cannot be established with confidence. As a result, traditional root cause analysis does not apply to nontransparent systems or applies only in a limited manner.

*d) Representative populations are observable.*

Successful statistical modeling depends on a sample size that is sufficient to achieve the desired statistical power and to support generalization to the target population. Precision and reliability improve as effective degrees of freedom increase. The required sample size increases with the number of parameters needed to represent the population because each parameter consumes degrees of freedom.

$$DF = N - P \quad [F4]$$

- $DF$  = Degrees of Freedom.
- $N$  = Sample Size.
- $P$  = Number of Parameters Being Estimated.

In a complex system, the parameter space spans the complete system and its interactions with the operating environment. When large samples are infeasible, traditional reliability engineering mitigates scale by decomposing the system into subsystems with distinct failure modes, enabling smaller test populations, reduced parameter counts, and reuse of data across subsystems. In contrast, AI models, including ensembles and hierarchical architectures, do not decompose cleanly for reliability purposes. Emergent behavior arises from interactions such that the whole exhibits properties that are not captured by the parts.

As a result, AI elements require an open-system treatment, and the number of relevant parameters expands with interacting components, configurations, and environmental states, rendering the sample sizes demand impractical by traditional techniques.

*2) Statistical Assumptions*

*a) Independent and identically distributed data.*

Complex systems are path dependent and stateful; failure processes depend on prior states, interactions, and environmental history. As a result, failure events exhibit serial dependence and heterogeneity across operating regimes. These elements are not independent or identically distributed; therefore, the assumptions of independence and identical distribution are not applicable to complex systems, such as AI-enabled safety-related systems.

*b) Negligible impact of outliers in the tails.*

Complex systems frequently exhibit fat-tailed behavior in which rare operating conditions and tail dependencies govern system risk. Emergent failure modes and mechanisms arise in these low-probability regimes, so treating tail observations as noise biases estimates and obscures dominant contributors to failure. The assumption that outliers have negligible impact does not hold for complex systems, including AI-enabled safety-related systems.

*c) Negligible change in variation over time.*

Complex systems co-evolve. Interactions among elements and with the environment alter operating regimes and error distributions in a recursive manner. Variance and higher-order moments drift over time, which violates the assumption of constant variability. In AI-enabled safety-related systems, changes in data, configurations, and environmental conditions introduce nonstationary, so models that presume fixed variation do not hold.

*d) Negligible complex interaction among variables.*

Complex systems exhibit high interactivity. Higher-order dependencies, second-, third-, and nth-order effects, can dominate behavior in rare operating regimes that precipitate emergent failures. Interaction terms are not small or strictly additive and coupling across subsystems produces nonlinear responses. Therefore, the assumption of negligible complex interaction among variables does not hold, particularly for AI-enabled safety-related systems.

*e) Example*

This section provides an example of improper application of traditional reliability engineering methods to complex systems, specifically AI-enabled safety-related systems.

Developers sometimes ask: “If the target is  $10^{-6}$  failures per hour, why is it not sufficient to operate for 1,000,000 hours without a failure?” The answer follows directly from the preceding analysis: the assumptions that underlie traditional methods, independent and identically distributed failures, constant failure rates, stationarity, limited interaction effects, and full explainability, do not hold for complex systems. As a result, absence of observed failures of a test interval does not validly establish the required failure rates for an AI-enabled safety-related system.

[Table 4] — Example Exercise Rationale

No	Description	Rationale
A1	Failure modes may be identified a priori.	The conditions that elicit emergent edge-case failure modes in an AI-enabled safety-related system cannot be specified a priori at the test-planning stage.
A2	Failure mechanisms are well characterized and explainable.	Observed test failures cannot be reliably attributed to specific internal mechanisms of the AI-enabled safety-related system.
A3	Failure behavior is explainable by known mechanisms.	Predictive claims regarding future failure behavior are valid only under the tested conditions; extrapolation beyond those conditions lacks empirical support.
A4	Representative populations are empirically observable.	A dataset comprising 1 million hours of failure-free operation provides little predictive value for field performance under open-world conditions and nonstationarity.
B1	The data is assumed to be independent and identically distributed.	Learning, updates, and stateful operation induce path dependence that violates the independent and identically distributed assumption.
B2	Outliers in the tails have negligible impact.	Outliers not observed during the 1-million-hour test may disproportionately influence failure behavior in deployment, particularly under heavy-tailed distributions.
B3	The data exhibits approximately constant variance over time.	Model updates and operational-environment shifts induce distribution shift and heteroscedasticity that are not captured by pre-deployment testing.
B4	Higher-order interaction effects among variables are negligible.	A dataset comprising 1-million hours of failure-free operation does not provide sufficient degrees of freedom for statistically significant reliability claims when higher-order interactions inflate the effective parameter space.

## V. A NOVEL FRAMEWORK FOR QUANTIFYING THE RELIABILITY OF AI-ENABLED SAFETY-RELATED SYSTEMS

This section applies principles from complex systems theory to address the previously identified limitations and establishes a

framework for quantifying the reliability of AI-enabled safety-related systems. The framework aligns evidence with properties of complex systems and provides the structure used in the analysis that follows.

### 1) Unsafe Hypothesis

Traditional functional safety validates safe behavior by explaining failure modes and mechanisms and by demonstrating prevention/mitigation. Testing in that paradigm largely confirms a safety hypothesis. This conformity mindset is at odds with the scientific method, which requires tests designed to falsify a hypothesis; a hypothesis that withstands progressively more severe falsification attempts earns practical credibility for engineering decisions.

For complex systems such as AI-enabled safety-related systems, a scientific-method-based approach is adopted:

*Null Hypothesis ( $H_0$ ) = The AI-enabled safety-related system is unsafe.*

[F5]

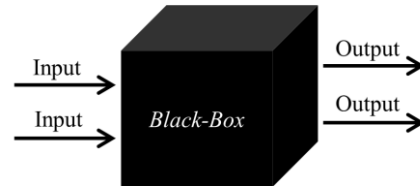
Tests are therefore designed to elicit and observe unsafe behavior. Although counterintuitive, systematically identifying the boundaries where unsafe emergent behavior occurs enables specification of operating regions in which the system demonstrates safe behavior with defined confidence.

### 2) Black-Box Treatment

Traditional functional safety for simple or complicated systems presumes explainability and transparency. In practice, modern software-based safety functions often operate beyond the limits where full explainability is feasible. Nevertheless, assurance processes frequently proceed as if complete explainability were still attainable.

Treating an AI-enabled safety-related system as a black box acknowledges these limits. Tests intentionally probe edge-case conditions to elicit failures and map operational boundaries. The resulting input–output evidence provides practical, operational explainability of observed behavior, even when internal mechanisms remain opaque, and supports defining the conditions under which the system operates safely with defined confidence.

[Figure 4] – Black-Box with Inputs and Outputs



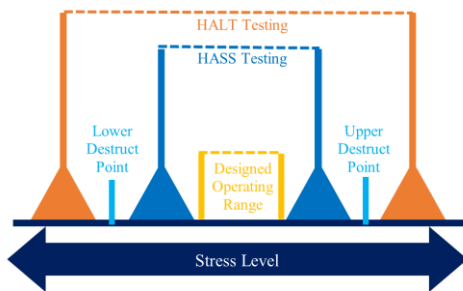
[12]

### 3) Testing First

Traditional functional safety for simple or complicated systems places analysis before testing: theoretical failure modes and mechanisms are identified a priori and then validated a posteriori via functional and fault-insertion tests.

Reversing the order by placing testing first enables the design of tests of progressively increasing severity to elicit emergent failure behavior. This approach parallels highly accelerated life testing (HALT) and highly accelerated stress screening (HASS) in hardware, which intentionally induce or replicate failure modes associated with manufacturing variation. The resulting empirical evidence then drives analysis, supports definition of operational boundaries, and informs risk controls for AI-enabled safety-related systems.

[Figure 5] – Diagram of HALT or HASS Testing



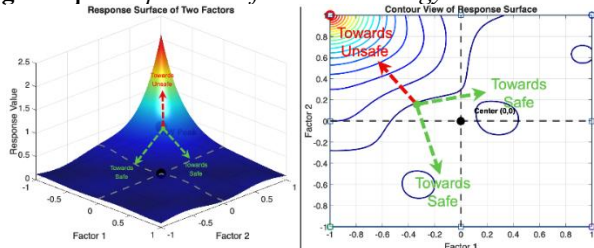
[13]

### 4) Edge Finding

Traditional functional safety for simple or complicated systems prescribes specific, constrained test conditions based on accepted procedures, focusing on conditions known to elicit failures and on validating that diagnostics or other means prevent dangerous undetected failures.

For complex systems such as AI-enabled safety-related systems, this approach expands to efficiently explore combinations of conditions, including rare and interacting factors. The objective is to locate operational boundaries, characterize pass-fail regions, and map conditions to observed behavior. Established design-of-experiments techniques, including Response Surface Methodology (RSM), support systematic exploration of the input space and identification of edge cases that precipitate emergent failures.

[Figure 6] – Response Surface Methodology



[22]

Once bounded operational regions of stable, safe behavior are identified through physical or virtual testing, and key controls are in place (fixed model version and configuration, controlled operating profile and environment, and trial resets that limit path dependence), the assumptions of traditional reliability engineering become locally tenable.

Constraining the domain reduces the effective parameter count  $P$  and addresses the degrees-of-freedom challenge  $DF = N - P$ . Instead of testing “the universe”, analysis focuses on “one room”, so the available observations  $N$  achieve usable statistical power. Within this defined region, traditional tools (e.g., Weibull analysis, FMEA, and FTA) can produce defensible estimates of failure rates and related reliability parameters.

### 5) Iterative Test and Analysis

Traditional functional safety for simple or complicated systems assumes a linear sequence of design, development, testing, and deployment. In software-based safety functions, this sequence can introduce months or years between releases.

For complex systems such as AI-enabled safety-related systems, especially those that learn or update regularly, an iterative test-and-analysis approach is required. The same unsafe-hypothesis framing, black-box treatment, testing-first posture, and edge-finding methods apply in each cycle to elicit emergent failures, refine operational boundaries, and confirm safe behavior. Executed rapidly, these cycles support near-real-time updates while maintaining documented evidence for safety claims and operational confidence.

## VI. DISCUSSION

Distinguishing complicated systems from complex systems is essential for quantifying reliability in AI-enabled safety-related systems. The foundational assumptions of traditional reliability engineering do not hold in general for complex systems. However, when the framework presented here is applied, falsification-oriented hypothesis testing, black-box treatment, testing-first practice, and systematic edge finding using methods such as Response Surface Methodology, bounded operational regions with stable behavior can be established. Within these regions, key controls fix configurations, constrain operating profiles, and limit path dependence, which reduces the effective parameter space. Under these localized conditions, the assumptions of traditional reliability engineering become materially valid, enabling the defensible application of conventional tools to estimate failure rates and related reliability parameters.

## VII. CONCLUSION

Development of AI-enabled safety-related systems requires explicit treatment of emergent behavior characteristic of complex systems. Observation of failure-free samples alone does not establish reliability and does not resolve limited explainability and transparency. Applying a falsification-oriented test strategy, black-box treatment, and systematic edge

finding defines bounded operating regions with stable behavior; within these regions, traditional reliability methods provide defensible estimates of failure rates and related parameters.

#### VIII.ACKNOWLEDGEMENT

The authors have no acknowledgments to make at time of publication.

## IX. REFERENCES AND BIBLIOGRAPHY

- [1] Exida, "Exida White Paper Library," November 2017. [Online]. Available: <https://www.exida.com/articles/Three-Barriers.pdf>. IEC 60050, "Artificial Intelligence," 29 March 2019. [Online]. Available: <https://www.electropedia.org/iev/iev.nsf/display?openform&ievref=171-09-17>.
- [2] International Electrotechnical Commission (IEC), "IEC 61508," July 2010. [Online]. Available: <https://www.iec.ch/functional-safety>.
- [3] ISO, "About ISO," ISO, 2024. <https://www.iso.org/about>
- [4] J.P. D. T. O'Connor and A. Kleyner, Practical Reliability Engineering. Chichester, UK: John Wiley & Sons Ltd, 2011. doi: <https://doi.org/10.1002/9781119961260>.
- [5] N. N. Taleb, "Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications," arXiv, 2nd rev. ed., arXiv:2001.10488 [stat.OT], 2022. doi: 10.48550/arXiv.2001.10488. [Online]. Available: <https://arxiv.org/abs/2001.10488>
- [6] "FMD-91 Failure Mode/Mechanism Distributions Reliability Analysis Center Reproduced from Best Available Copy." Accessed: Sep. 17, 2025. [Online]. Available: <https://apps.dtic.mil/sti/tr/pdf/ADA259655.pdf>
- [7] Els-cdn.com, 2025. <https://ars.els-cdn.com/content/image/1-s2.0-S1384107618302458-gr3.jpg> (accessed Sep. 17, 2025).
- [8] "IEC 60812:2018," Webstore.iec.ch, 2018. <https://webstore.iec.ch/en/publication/26359>
- [9] R. Hanel, S. Thurner, and P. Klimek, Introduction to the Theory of Complex Systems. 2018. doi: <https://doi.org/10.1093/oso/9780198821939.001.0001>.
- [10] "IEC 61496-1:2020," Webstore.iec.ch, 2020. <https://webstore.iec.ch/en/publication/63115>
- [11] A. Weis, Black-Box with Inputs and Outputs. 2025.
- [12] A. Weis, Diagram of HALT or HASS Testing. 2025.
- [13] "IEC 61508:2010 CMV," Webstore.iec.ch, 2020. <https://webstore.iec.ch/en/publication/22273>
- [14] "IEC 60050 - International Electrotechnical Vocabulary - Details for IEC number 351-57-06: 'functional safety,'" Electropedia.org, 2025. <https://www.electropedia.org/iev/iev.nsf/display?openform&ievref=351-57-06> (accessed Sep. 17, 2025).
- [15] "IEC 60050 - International Electrotechnical Vocabulary - Details for IEC number 171-09-17: 'artificial intelligence,'" Electropedia.org, 2019. <https://www.electropedia.org/iev/iev.nsf/display?openform&ievref=171-09-17> (accessed Sep. 17, 2025).
- [16] Iso.org, 2025. <https://www.iso.org/obp/ui#iso:std:iso:22291:ed-1:v1:en:term:3.11> (accessed Sep. 17, 2025).
- [17] "IEC 60050 - International Electrotechnical Vocabulary - Details for IEC number 428-04-23: 'diagnostic coverage,'" Electropedia.org, 2021. <https://www.electropedia.org/iev/iev.nsf/display?openform&ievref=428-04-23> (accessed Sep. 17, 2025).
- [18] "IEC 60050 - International Electrotechnical Vocabulary - Details for IEC number 428-04-33: 'average frequency of a dangerous failure per hour,'" Electropedia.org, 2024. <https://www.electropedia.org/iev/iev.nsf/display?openform&ievref=428-04-33> (accessed Sep. 17, 2025).
- [19] Iso.org, 2025. <https://www.iso.org/obp/ui#iso:std:iso:tr:12489:ed-1:v1:en:term:3.1.16> (accessed Sep. 17, 2025).
- [20] "IEC 60050 - International Electrotechnical Vocabulary - Details for IEC number 428-04-11: 'safe failure fraction,'" Electropedia.org, 2021. <https://www.electropedia.org/iev/iev.nsf/display?openform&ievref=428-04-11>
- [21] D. Beam, Response Surface with Unsafe and Safe Regions. 2025. "IEC 61508-6:2010", Webstore.iec.ch, 2019. <https://webstore.iec.ch/en/publication/5520>
- [22] A. Gautam, Hardware Component Weibull Plot. 2025.

## X. TABLES

- [Table 1] Example of Failure Mode Distribution (FMD-91)
- [Table 2] Reliability Engineering Epistemic Assumptions
- [Table 3] Reliability Engineering Statistical Assumptions
- [Table 4] Example Exercise Rationale

## XI. FIGURES

- [Figure 1] Hardware Component Weibull Plot
- [Figure 2] Hardware System Fault Tree Analysis
- [Figure 3] Three Body Problem
- [Figure 4] Black-Box with Inputs and Outputs
- [Figure 5] Diagram of HALT or HASS Testing
- [Figure 6] Response Surface with Unsafe and Safe Regions

## XII. FORMULAS

- [F1] Failure Distribution Function
- [F2] Reliability Function
- [F3] Hazard Function
- [F4] Degrees of Freedom
- [F5] Null Hypothesis