Becoming Bayesian*

Making Decisions with Bayes Factors in Flight Test

Brig Gen Douglas "Beaker" Wickert, PhD[1]
Air Force Materiel Command

The fundamental cause of trouble in the modern world is that the stupid are cocksure while the intelligent are full of doubt

Bertrand Russell

Decision-making in flight test is done under a persistent state of uncertainty. The risk management frameworks used in flight test are typically probability-based, but assessing probability is unreliable due to the epistemic uncertainty inherent in flight test. Bayesian reasoning, a method that continuously updates our beliefs about the world in light of new evidence, is widely used in scientific fields. This paper describes an approach using *Bayes Factors* to update our prior judgment on whether a proposed test is safe or unsafe. The use of Bayes Factors also provides a systematic structural framework with a common lexicon, leveraging the collective wisdom of the crowd during test and safety planning, to enable meaningful conversations about risk. Multiple examples of the practical use of Bayes Factors are provided. Lessons learned and recommendations for flight test risk management are offered.

Key Words: Bayes Factor, Risk, Risk Management, Flight Test, Uncertainty

Nomenclature

BF	Bayes Factor (likelihood ratio)
$\Theta^{(0)}$	prior odds ratio
$\Theta^{(1)}$	posterior odds ratio
H	hypothesis
$\neg H$, H^c	negation or complementary (opposite) hypothesis
D	data
$\mathrm{P}(\mathrm{H} \mid D)$	conditional probability of hypothesis given the data

* DRAFT v2.1

[1] Associate Fellow SETP

I. Uncertainty & Flight Test

Uncertainty is unavoidable and ubiquitous. A fact more true in flight test than in many other professions. To reduce uncertainty about a system under test, flight testing deliberately probes the unknown, "pushing the edge of the envelope," identifying deviations in system performance from the intended design. Test also reveals unexpected behavior, model flaws, system deficiencies, poor human factor designs, and inaccurate design assumptions. Due to that inherent uncertainty, flight test is risky and potentially dangerous. A century of flight test has led to the development of a robust approach to risk management and a systematic, deliberate way of approaching the unknown—*Ad Inexplorata* was adopted in 1953 as the motto of the Air Force Flight Test Center. The "X" of the Society of Experimental Test Pilots, established in 1955, stands for the *unknown*.

Flight testers are, by virtue, professional risk managers. The Air Force Test Center[2] has ventured into the unknown for more than 80 years at Edwards Air Force Base. Over the last several years, a new approach to thinking about uncertainty has emerged at Edwards. Built around the Bayesian reasoning techniques that are now common in scientific data analysis, the approach has proven to be a useful framework for thinking about risk.

In many ways, this paper is the next step in a career spent thinking about risk management and uncertainty. In a 2018 SETP paper, the author attempted to describe the attributes of a risk-aware culture and techniques for cultivating what we called ***Risk Awareness*** [1]. Risk Awareness was defined in an analogy to Situational Awareness: the *perception of the elements of uncertainty and the potential, projected outcomes resulting from uncertainty.*[3] In contrast to the MIL-STD-882E definition of risk, "the combination of the severity of a mishap and the probability of that mishap occurring" [3], the author deliberately avoided considering 'probability-of-occurrence' in the definition of Risk Awareness because of the inherent unreliability of probabilities. This is particularly true given the epistemic uncertainties present in flight test. Since 2018, the author has "*Become Bayesian*" and has embraced Bayes Factors as a way of dealing with the inherent difficulty in assessing probabilities. This paper describes the approach.

In his reflection on a "whole career working on investigations aimed at reducing uncertainty about what is happening, what might happen, and even the reasons why things happen," the eminent statistician Sir David Spiegelhalter, Professor Emeritus at Oxford and past President of the Royal Statistical Society, writes that: "once we accept a personal, subjective view of probability and uncertainty, we are led naturally to *Bayesian analysis*, in which we use the theory of probability to revise our beliefs in the light of new evidence" [4].

In simple terms, a Bayesian approach is one in which we continuously update our beliefs about the world in the light of new evidence. A little more formally and in the terms used in Bayesian inference:

[2] the Air Force Flight Test Center (AFFTC) was redesignated as the Air Force Test Center (AFTC) in 2012 as part of a major reorganization of Air Force Materiel Command; the reorganization brought AFFTC, the Arnold Engineering Development Center, and the Air Armament Center together into a single Center

[3] situational awareness is defined as the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future [2]

$$\begin{pmatrix} Posterior \\ Belief \end{pmatrix} \propto \begin{pmatrix} Likelihood \\ Ratio \end{pmatrix} \times \begin{pmatrix} Prior \\ Belief \end{pmatrix} \tag{1}$$

This paper describes an approach of using Bayes Factors (a likelihood ratio given a set of evidence) to update our prior judgment on whether a proposed test is safe or unsafe. The evidence may be engineering predictions, test data, unexpected test events, test team training and readiness, concerns of programmatic drift, *etc.* The use of Bayes Factors also provides a structural framework with a common lexicon during test and safety planning to facilitate meaningful conversations about risk. This can help leverage the collective "wisdom of the crowd" during test planning and execution.

Before describing the formal technique of using Bayes Factors in Section III, we begin with a description of probability, judgment, and decision-making under uncertainty (Section II). Section II may be skipped without any loss in understanding the application of Bayes Factors, but the description of a century-plus-long intellectual journey through uncertainty and decision-making contributes to understanding the theoretical foundation of our recommendations for making good decisions. After describing the Bayes Factor approach, we present eight applications of the use of Bayes Factors using both test and non-test examples (Section IV). Each example illustrates a different attribute and lesson learned from "*Becoming Bayesian*," which are discussed and summarized in Section V. Two appendices provide the mathematical basis for the derivation of Bayes Factors from Bayes' Theorem (Appendix A) and of the use of the Beta and Gamma distributions in Bayesian reasoning (Appendix B).

II. Risk & Decision-Making Under Uncertainty

Decision-making under uncertainty is a fundamental challenge encountered in many human endeavors, from medical diagnosis and financial investment to military strategy and engineering design. A rich body of literature exists on the subject [5, 6, 7]. Uncertainty is particularly ubiquitous and persistent in flight test and is often a topic at SETP and SFTE symposia.

The author's preferred definition of uncertainty comes from Sir David Spiegelhalter's excellent book on the topic: *uncertainty* is the *conscious awareness of ignorance* [4]. Our ignorance—our lack of knowledge of what "what we know" or "what we can know"—stems from two primary factors: the theoretical limits of knowledge (*epistemic* uncertainty) and inherent variability or randomness in some systems (*aleatory* uncertainty).

Risk is a fundamental consequence of uncertainty. The International Organization for Standardization (ISO) defines risk as "the effect of uncertainty on objectives" [8], while the Air Force defines it as "probability and severity of loss or adverse impact from exposure to various hazards" [9]. The definition used by the Air Force Test Center parallels the Air Force definition: test risk is the "combination of the severity of the mishap and the probability that the mishap will occur" [10].

All definitions of risk effectively have three elements in common: 1) scenarios; 2) consequences; and 3) likelihoods. ISO's definition of risk management is the "coordinated activities to direct and control an organization with regard to risk." Similarly, the Air Force defines risk

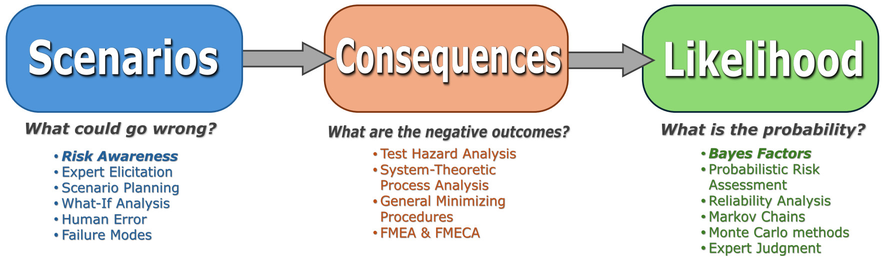


Figure 1: *Risk Management Framework and Associated Analysis Tools*

management as a "decision-making process to systematically evaluate possible courses of action, identify risks and benefits, and determine the best course of action" [11]. Risk management is equivalent to answering three questions related to the three elements in Figure 1:

1. What can go wrong? (*generally an undesired/unexpected event*)
2. What are the resulting consequences? (*negative outcomes from the unexpected event*)
3. What is our expectation of the likelihood? (*probability of occurrence*)

The author has long been uncomfortable with the third question. This is primarily due to the difficulty in making probability assessments with any confidence (Section II.A). The emphasis in the Risk Awareness framework [1] on uncertainty was an attempt to avoid the question of likelihood. However, as discussed in Section II.C, we implicitly make probability assessments every time we make decisions under uncertainty. Thus, probability is unavoidable and integral to risk management.

Within the three-element mental model of risk in Figure 1, Risk Awareness supports, guides, and informs the scenario analysis and planning process. Test Hazard Analysis (THA), Failure Mode and Effects Analysis (FMEA), System-Theoretic Process Analysis (STPA), and other system engineering tools are useful in defining the scope and impact of consequences. Many tools exist to support the third element: Monte Carlo methods, Probability Risk Assessment [12], and expert opinions are examples of approaches to assess likelihoods. In this paper, we offer Bayes Factors as another. Our adoption and approach to using Bayes Factors to inform risk management is motivated by two axioms that serve as guiding principles:

1. Probability is unreliable (probability does not exist, but it is useful to act as if it does)
2. Wisdom of the crowd (the fact that aggregate judgments can often outperform individual assessments by experts)

A. Probability is Unreliable

Probability estimates are essential to risk management and decision-making under uncertainty. Given the unavoidable need to assess probability, it is worth noting that we do not even have a good way to define it. Bertrand Russell, whom we already quoted in the epigraph, remarked in a 1929 lecture that "Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means" [13].

There is a long and rich history of debate on the nature of probability. The major interpretations and attempts to define probability since Pascal and Fermat first corresponded[4] on the topic in the late 1600s have included: *classical* probability (Laplace), *logical/evidential* probability (Keynes, Carnap), *subjective* probability (Bayes, de Finetti), *frequency* interpretations (von Mises, Fisher, Neyman), *propensity* interpretations (Popper), and *axiomatic* definitions (Kolmogorov). Rehashing the metaphysical debate is not our current purpose.[5] But awareness of the difficulty to even define *probability* means that we should expect difficulty in assessing it.

The author finds the most useful statement on probability to be consistent with Italian mathematician Bruno de Finetti's perspective that "Probability does not exist" [15]. By this, de Finetti meant that probability is not an inherent property of the world, but rather a degree of belief held by an individual, revealed through their choices. This view was opposed by both *frequentists* like Fisher, Neyman, and Pearson, who saw probability as long-run frequency of expectations, and *objectivists* like Keynes, who saw probability as the objective logical relation between propositions [16]. Ramsey, whom we will meet again below, was sharply critical of Keynes and proposed a subjective interpretation in which probability reflects an individual's coherent degrees of belief [17]. This disagreement in the nature of probability would be entwined with the bitter debate that raged for much of the 20th century between *frequentists* and *Bayesians*.

Today, the philosophical and academic debate is over for all practical purposes, with most practitioners[6] adopting a Bayesian understanding that rests on de Finetti's subjective interpretation of probability [18, 4, 19, 20, 21, 22]. Probability is best understood as our subjective relationship with what we know about the world, and Bayesian reasoning is the process by which we update our subjective understanding. Probability is then a direct measure of our uncertainty. If I flip a coin and ask you about *your probability* of the heads/tails outcome, your expectation is unaffected by whether I look at the result after the flip. "*My probability*" given knowledge of the outcome is obviously different from *your probability* of heads/tails.[7]

Estimating probabilities in the real, ambiguous world is notoriously challenging. For example, what is the probability of interest rates being above or below 2.5% a year from now? Ten years from now? What is the probability that Chairman Xi decides to invade Taiwan? What is the probability that the Atlantic Meridional Overturning Circulation (AMOC)[8] collapses in the next twenty years? In the next fifty years?

Engineering predictions can also be notoriously unreliable. For example, reliability calculations by McDonnell Douglas had estimated that the probability of a DC-10 experiencing a

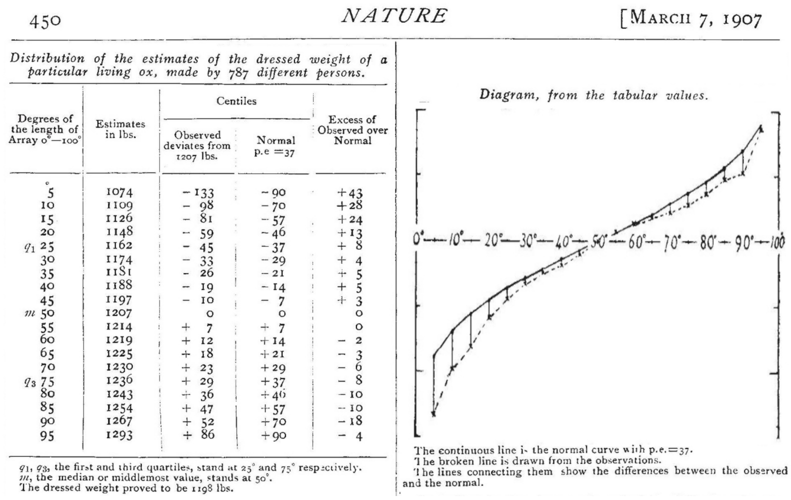
[4] a collaboration sparked by a problem presented to Pascal by Parisian gambler and writer, the Chevalier de Méré, who wanted to know how to divide the stakes of a prematurely interrupted game fairly

[5] there are many excellent summaries of the debate; see [14] for example

[6] in addition to being standard in data analysis, Bayesian techniques are widely used in AI and machine learning; large-language models are typically built on Bayesian foundations

[7] this example also illustrates the distinction between aleatory and epistemic uncertainty: before the flip, there is aleatory uncertainty of the outcome; after the flip, my knowledge of the outcome does not affect your residual epistemic uncertainty of the outcome

[8] AMOC is the "conveyor belt" moving warm, salty water from the tropics northward and colder, deeper water southward; the circulation is a crucial component of the global climate system


450 *NATURE* [MARCH 7, 1907

Distribution of the estimates of the dressed weight of a particular living ox, made by 787 different persons.

Degrees of the length of Array 0°—100°	Estimates in lbs.	Centiles		Excess of Observed over Normal
		Observed deviates from 1207 lbs.	Normal p.e =37	
5°	1074	−133	−90	+43
10	1109	−98	−70	+28
15	1126	−81	−57	+24
20	1148	−59	−46	+13
q_1 25	1162	−45	−37	+8
30	1174	−33	−29	+4
35	1181	−26	−21	+5
40	1188	−19	−14	+5
45	1197	−10	−7	+3
m 50	1207	0	0	0
55	1214	+7	+7	0
60	1219	+12	+14	−2
65	1225	+18	+21	−3
70	1230	+23	+29	−6
q_3 75	1236	+29	+37	−8
80	1243	+36	+46	−10
85	1254	+47	+57	−10
90	1267	+52	+70	−18
95	1293	+86	+90	−4

q_1, q_3, the first and third quartiles, stand at 25° and 75° respectively.
m, the median or middlemost value, stands at 50°.
The dressed weight proved to be 1198 lbs.

Diagram, from the tabular values.

The continuous line is the normal curve with p.e.=37.
The broken line is drawn from the observations.
The lines connecting them show the differences between the observed and the normal.

Figure 2: *Figures from Sir Francis Galton's 1907 Nature article noting the wisdom of crowds*

simultaneous loss of engine thrust and asymmetric leading edge slats to be less than one-in-a-billion (10^{-9}) [23]. But this identical combination of failures occurred four times in four years and was the root cause of the 1979 crash at Chicago O'Hare, still the deadliest single airline accident on US soil. Similarly, various probability estimates by NASA engineers for loss of the Space Shuttle ranged from 1-in-10 to 1-in-100,000, a discrepancy of four orders of magnitude [24]. More recently, the flight testing of a new platform at Edwards AFB experienced a 10^{-9} anomaly during the first 10 hours of test operations.[9]

Probability is unreliable and 'probably' does not objectively exist, yet we need it for decision-making under uncertainty, so it is useful to act as if it does exist. Bayes Factors offer one mechanism of coming to terms with the inherent unreliability of probability. And, as we shall see, Bayes Factors provide a framework that can leverage the wisdom of the crowd, improving our collective assessment of probability and our decision-making.

B. Wisdom of the crowd

The "wisdom of the crowd" is the phenomenon where aggregated judgments from diverse groups outperform individual assessments by experts. The principle was first quantitatively observed by Sir Francis Galton [25] at a 1906 country fair in Plymouth, England. When Galton analyzed nearly 800 guesses from fairgoers attempting to estimate the weight of an ox, he discovered that while individual estimates varied widely, the median of all guesses was remarkably accurate—within one percent of the actual weight of 1,198 pounds (Figure 2). This result has been repeatedly validated: collective intelligence often emerges from the aggregation of diverse, independent judgments, even when individual participants lack specific expertise.

A century after Galton, Philip Tetlock's Good Judgment Project provided rigorous empirical validation of the principle of the wisdom of the crowd through large-scale forecasting tournaments in

[9] if the aircraft flew for 8 hours a day, 365 days a year, the mean time for the first such expected failure should be 100 years

which teams of ordinary citizens, when properly structured and trained in probabilistic reasoning, outperformed intelligence analysts with access to classified information [26]. Tetlock identifies the conditions for effective crowd wisdom: diversity of perspectives, independence of initial judgments, and an aggregation mechanism. Thru the Good Judgment Project, Tetlock and journalist Dan Gardner [27] claim to have also identified "superforecasters" who consistently outperform both crowds and experts. Fittingly to our present purpose, one of the common attributes of superforecasters' success was "being Bayesian," *i.e.*, regularly updating their beliefs in response to new information.[10]

In the "The Wisdom of Crowds" Surowiecki describes four conditions necessary for crowd wisdom to emerge: diversity of opinion among participants, independence of individual judgments (avoiding cascading influence), decentralization that allows people to draw on local knowledge, and an effective aggregation mechanism to combine individual inputs into collective decisions [29]. These are also worthwhile guides for any attempt to structure our flight test safety review processes and harness the collective wisdom of the crowd.

In the author's experience, the flight test community does a good job of considering and synthesizing judgments from diverse teams, as pilots, engineers, maintainers, flight doctors, and independent safety reviewers typically participate in comprehensive safety reviews. If improvement is needed, it is in having a framework and common lexicon for meaningful conversations about risk. Without a common lexicon, it is difficult for experts from different fields to communicate and weigh relative likelihoods. This is one of our motivations for encouraging a widespread adoption of Bayes Factors.

C. Making Decisions Under Uncertainty

Managing risk is essentially the same as making decisions under uncertainty, and in making decisions under uncertainty, we are either implicitly or explicitly making probability judgments. The psychology of decision-making is a deeply researched field. An excellent recent paper surveys the field from the Greeks through 21st-century thinkers, including Kahneman and Gigerenzer [30]. Risk Awareness [1] discussed in detail the complementary perspectives of decision-making under uncertainty offered by Kahneman [31] and Gigerenzer [32]. Here, we provide a brief overview of the major 20th-century intellectual perspectives on risk and decision-making.

[10] other attributes include good qualities that all testers wanting to improve decision-making should aspire to:

- *Bayesian*: regularly updating beliefs in response to new information
- *Aggregation*: systematically using multiple sources of information
- *Openness*: openness to new knowledge and actively seeking contrary views and perspectives
- *Meta-cognition*: insight into their own thinking and awareness of cognitive biases
- *Decomposition*: breaking complex problems into smaller components more tractable to analysis
- *Humility*: acknowledging uncertainty, expressing predictions probabilistically rather than definitively, and readily admitting errors
- *Hedgehog vs. Fox*: from Isaiah Berlin's 1953 essay: "the fox knows many things, but the hedgehog knows one big thing" [28]; describes a fundamental difference in thinking styles—*foxes* are generalists with diverse interests and strategies, whereas *hedgehogs* are specialists who relate everything to a single idea or system

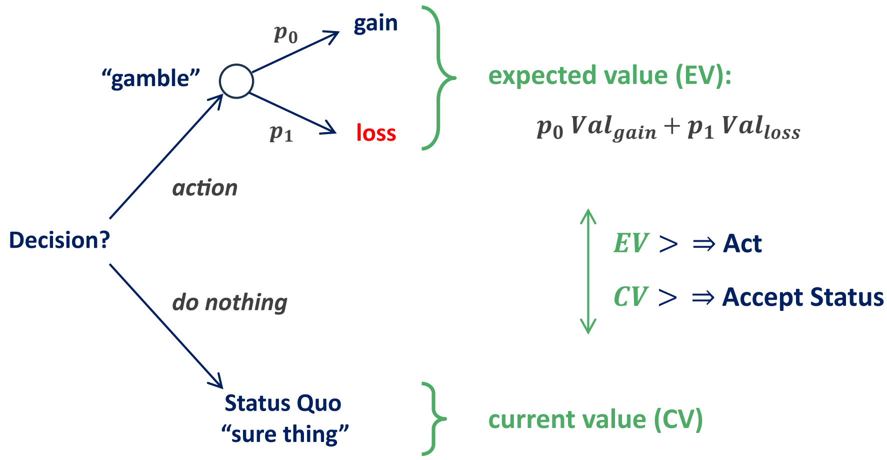


Figure 3: *Ramsey's Decision Theory Calculus*

Knight (1921):[11] first articulated the distinction between *risk*—situations where the outcomes are unknown but governed by probability distributions—and *uncertainty*—situations in which outcomes and probability models are unknown [33]. In modern terms, we refer to aleatory and epistemic uncertainty. The Risk Awareness framework [1] argued that different cognitive and risk management tools were needed for the different risk and uncertainty domains. Stirling made a similar argument for policy considerations in an influential letter in Nature [34].

Ramsey (1926):[12] Ramsey's decision calculus formalized the idea that rational agents choose actions that maximize their expected utility, where utility represents the subjective value of outcomes and probabilities represent degrees of belief (Figure 3) [17, 35]. This framework provides a mathematical foundation for treating uncertainty as a quantifiable factor in decision-making and establishes a theoretical basis for our Bayes Factor approach to risk management.

Simon (1947):[13] introduced *bounded rationality* and argued that human decision-makers, faced with cognitive limitations, incomplete information, and time constraints, engage in *satisficing* behavior—searching through alternatives until finding one that meets their minimum acceptable criteria (Figure 4) [36, 7]. Decision makers often stop searching when they find an option that *satisfices* some objective criteria. There is an implicit, subconscious probability assessment of the likely outcome when one stops searching through the space of bounded rationality.

Kahneman and Tversky (1974):[14,15] in the highly influential paper, "Judgment Under Uncertainty: Heuristics and Biases," they highlighted three types of heuristics by which probabilities are assessed and judgments made: availability, representativeness, and anchoring and adjustment [5]. Kahneman spent his career systematically revealing the flaws in human reasoning and introduced the dual-system framework that is now popular and well-known:

[11] Frank Knight (1885–1972): American economist and one of the founders of the Chicago School

[12] Frank P. Ramsey (1903–1930): a brilliant British mathematician, philosopher, and economist whose tragically brief career left foundational contributions to decision theory that remain central to modern risk analysis

[13] Herbert Simon (1916–2001): a Nobel Prize-winning American economist and cognitive scientist

[14] Daniel Kahneman (1934–2024): Nobel-prize winning Israeli-American psychologist; collaborated with Tversky in the development of prospect theory and established the field of behavioral economics

[15] Amos Tversky (1937–1996): Israeli mathematical psychologist; collaborated with Kahneman in the development of prospect theory and the field of behavioral economics

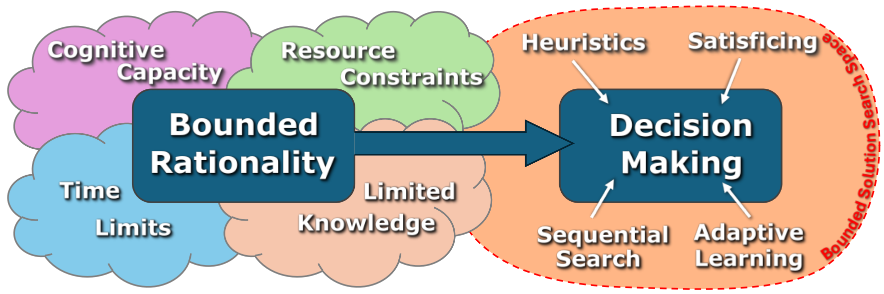


Figure 4: *Simon's Bounded Rationality*

- System I: fast and intuitive, but prone to cognitive biases and heuristic shortcuts that often lead to demonstrably poor decisions
- System II: slow and deliberate, employing careful analysis to overcome these biases and make rational conclusions

Gigerenzer (1996):[16] building on Simon's bounded rationality, Gigerenzer argues that fast-and-frugal heuristics—simple rules of thumb—are often ecologically rational and has demonstrated that heuristics can outperform complex analytical methods in environments characterized by high uncertainty, sparse data, and time pressure [37, 38].

The Risk Awareness framework [1] emphasized the distinction between Kahneman and Gigerenzer's work as particularly relevant in flight test. Risk Awareness argued that the domain of uncertainty should determine the most appropriate cognitive approach. System II analytical thinking proves valuable in the risk domain where probabilities are well characterized, while experienced heuristics and "gut feelings" often provide better guidance in the pure uncertainty domain where many critical flight test decisions are made.

Klein (1999):[17] Klein built his Recognition-Primed Decision (RPD) Model on the basis of observations of experienced professionals, *e.g.*, firefighters and military officers, who make rapid and effective decisions without, from his view, explicitly comparing multiple options [39]. RPD involves a two-stage process: an intuitive recognition of a *plausible* course of action based on past experiences (pattern matching) and a mental simulation of that action to ensure its *viability*. Both *plausibility* and *viability* require an implicit assessment of the likelihood of success, a probability assessment. Klein's RPD model and his description of naturalistic decision-making influenced the Army Field Manuals on Mission Command and the Army's formal curriculum on intuitive decision-making [40].

The intent of this survey of 20th-century research into decision-making is to underscore that, regardless of model—rational utility, bounded rationality models, slow thinking, fast-and-frugal decisions—decision-making under uncertainty requires an assessment of the expected probability of success (or failure) of an outcome. The better and more self-consistent our probability assessments are, the better risk-informed decisions we will make. We need a systematic method for weighing potential outcomes against their likelihood. We should also employ a common lexicon to harness the wisdom of the crowd to improve the accuracy of our collective judgments. This is the motivation

[16] Gerd Gigerenzer (1947-): German psychologist
[17] Gary Klein (1944-): American research psychologist who worked a U.S. Air Force psychologist in the mid-1970s

for encouraging the adoption of Bayes Factors in flight test.

III. Bayes Factors

Our knowledge of the world evolves with new evidence and data over time. We are, almost by nature, "born Bayesians."[18] The Bayes Factor approach recommended here provides a mathematical framework for combining objective and subjective measures of expectation and for updating our beliefs in light of new evidence and data. This makes it particularly well-suited to the iterative nature of flight test where our understanding evolves continuously from engineering design reviews through bench tests and build-up test campaigns.

The Bayes Factor was introduced by Jeffreys[19] in his 1939 book *Theory of Probability* [19]. With two complementary hypotheses, the Bayes Factor is the posterior odds of one hypothesis when the prior probabilities of the two hypotheses are equal. More simply, the Bayes Factor, or BF, is the likelihood ratio assessing the relative weight of evidence, data, or judgment between two competing hypotheses. Bayes Factors began to see more widespread adoption by statisticians, scientists, courts, and medical researchers following Kass and Raftery's 1995 paper on the subject [43].

A direct application of Bayes' Theorem (Equation A-3) to data is often difficult, because it requires an assessment of the probability of observing a particular piece of data, D (see the example in Section IV.B). Fortunately, the use of Bayes Factors (BF) avoids this step (Appendix A). The Bayes Factor for a particular piece of data or evidence is the likelihood ratio of observing that piece of data under two competing hypotheses. In our flight test context, this could be the hypothesis that a test is safe vs the complementary hypothesis that a test is unsafe. Thus, the Bayes Factor is

$$BF \equiv \frac{P(D|H_{safe})}{P(D|H_{unsafe})} \tag{2}$$

Any underlying error or unacknowledged assumption affecting the data or the accuracy of our judgment should be common to both the numerator and denominator in Equation 2 and will cancel out. This is one of the advantages of using Bayes Factors instead of the conditional probability.

To use the Bayes Factor (Equation 2) we have to express our probability in terms of odds ratios. Although common in gambling, expressing aleatory uncertainty in terms of odds is not as common as doing so in terms of probabilities. The odds that a die roll is a '6' are 1-to-5, 1:5, or 5:1 against. It is straightforward to convert probabilities into odds ratios, Θ, by

$$\Theta = \frac{p}{1-p} \tag{3}$$

Odds ratios are converted back into probabilities by

[18] Bayesian Brain Theory, a cognitive theory of the mind, proposes that the brain forms Bayesian predictive models of sensory inputs to enable adaptive behavior [41, 42]

[19] Sir Harold Jeffreys (1891–1989), a Fellow of the Royal Society, British geophysicist and statistician

$$p = \frac{\Theta}{\Theta + 1} \tag{4}$$

With the prior odds, $\Theta^{(0)}$, established, the posterior odds, $\Theta^{(1)}$, informed by the data are

$$\Theta^{(1)} = BF \cdot \Theta^{(0)} \tag{5}$$

Multiple pieces of data, D_i, can be considered simultaneously. The final result is the product of all the individual Bayes Factors for each piece of data, evidence, professional engineering judgment, *etc.*, so the final posterior odds are

$$\Theta^{(1)} = BF_1 \cdot BF_2 \ldots \cdot BF_n \cdot \Theta^{(0)} = \prod_i BF_i \cdot \Theta^{(0)} \tag{6}$$

New posterior odds using new Bayes Factors from new data can also be updated over time. This is helpful as additional build-up tests or other data about the system under test become available. The previous posterior odds simply become the new prior odds. Thus

$$\Theta^{(2)} = BF_{n+1} \cdot \Theta^{(1)} \tag{7}$$

Equation 4 is still used to convert the posterior odds back to a probability.

IV. Examples

We present four non-test and four flight test examples of using Bayes Factors to make judgments and conclusions. Each of the examples of Bayes Factors to decision-making under uncertainty was chosen to highlight different aspects of the application. We conclude each example with a summary of the lessons learned from the practical application and use of Bayes Factors.

A. How good is that Amazon Review?

Consider an Amazon product review—or hiring reference, restaurant or vacation recommendation, *etc.*—with only five individual reviews. Four are positive (*e.g.*, "five-star") reviews/recommendations and one is neutral or negative (*e.g.*, "two-star"). Given the data with an average 80% satisfaction level, what is your expectation of a favorable experience?

The Bayes Factor is the ratio of the probability of observing this data under the two hypotheses: $H_{favorable}$ and $H_{unfavorable}$. Under the hypothesis that you have a favorable experience, the probability of having observed the review data is 5/6, *i.e.*, there are now six reviews with five favorable outcomes. Likewise, under the hypothesis that you have an unfavorable experience, the probability of having observed the data is 2/6. Thus, the Bayes Factor is:

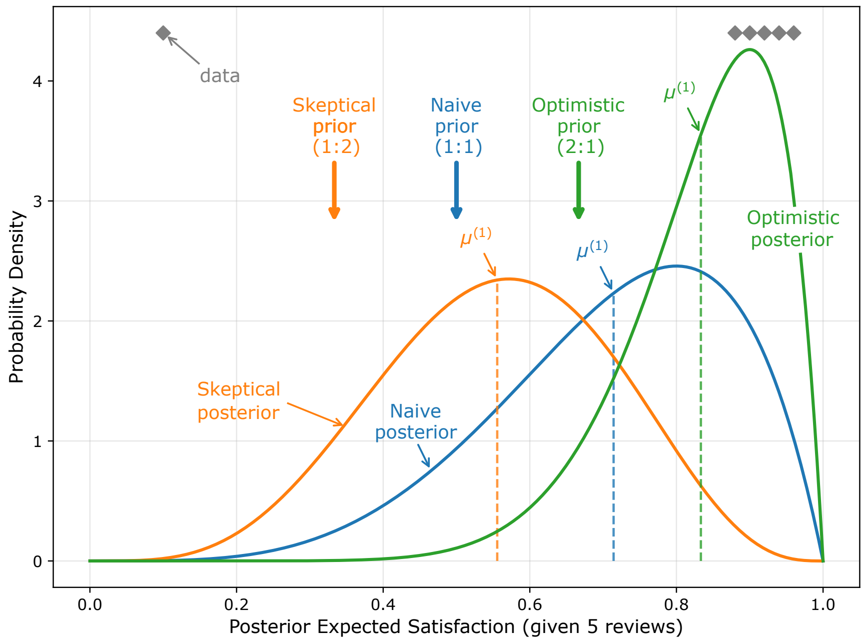


Figure 5: *Posterior Expected Satisfaction Given Prior Review Data*

$$BF = \frac{P(D|H_{favorable})}{P(D|H_{unfavorable})} = \frac{5/6}{2/6} = \frac{5}{2} \tag{8}$$

The posterior odds will depend on the prior odds. For “naive odds”, *i.e.*, there is no *a priori* reason to expect either a favorable or unfavorable outcome, the prior odds are 1:1. Thus

$$\Theta^{(1)} = BF \cdot \Theta^{(0)} = \left(\frac{5}{2}\right)\left(\frac{1}{1}\right) = \frac{5}{2} \tag{9}$$

Perhaps you generally have favorable experiences, with two favorable hiring decisions for every one that you regret. Then your prior odds might be 2:1. Alternatively, you may be a more critical skeptic than most reviewers (or hiring authorities) and assign your prior odds at 1:2. Table 1 summarizes the impact of these various priors. With naive odds, despite an 80% positive review history, you only have a 35% chance of having a favorable experience ($>$ 80% satisfaction). The combination of optmistic prior odds and 4/5 positive prior reviews gives you a 68% expectation of having a favorable experience. These somewhat low expectations are the consequence of the relatively sparse data. As shown in Appendix B, increasing the amount of data will ultimately overwhelm your prior odds and converge to the “true” distribution. Figure 5 shows the posterior expected distributions for each of the three priors in Table 1.

This example illustrates that our judgments and conclusions, particularly with sparse data, are heavily influenced by our prior expectations. Differences of opinions in conclusions are either due to different BFs, *i.e.*, the relative likelihood of the data under different hypotheses, or due to different prior odds. By framing the question in terms of Bayes Factors and articulating our numbers, we can expose those differences in a meaningful way. The use of BFs also forces us to consider

Table 1: Bayes Factors - Amazon Review

	$\Theta^{(0)}$	BF	$\Theta^{(1)}$	Mean	95% CI	$P(p > 0.8)$
Naive Prior	1:1	5/2	5:2	0.71	(0.245, 0.843)	35%
Optimistic Prior	2:1	5/2	5:1	0.83	(0.587, 0.977)	68%
Skeptical Prior	1:2	5/2	5:4	0.56	(0.245, 0.843)	6%

alternative hypotheses explicitly and is thus a good guard against confirmation bias.[20] The use of Bayes Factors will not eliminate confirmation bias, but acknowledging alternative hypotheses (and requiring a numerical assessment of them!) is a useful antidote. The surprisingly low expectation of a favorable experience, stemming from the sparse data in this example, serves as a cautionary example for avoiding the temptation for optimism bias and other planning fallacies.

Lesson Learned 1: Our judgments and conclusions are strongly influenced by our prior expectations
Lesson Learned 2: The BF-framework requires us to acknowledge alternate hypotheses explicitly (and may guard against confirmation bias, planning fallacy, blind spots, and other cognitive errors)

B. Kahneman and Tversky's Taxicab Problem

The "Taxicab Problem" is a classic probability puzzle that illustrates the base rate fallacy. Amos Tversky and Daniel Kahneman first offered this problem to study subjects in 1972 and published additional results in 1981 [44]. They concluded that people underweight or neglect base rate data leading to incorrect probability judgments. The "Taxicab Fallacy" is also frequently used as an example of a common cognitive shortcut known as the representativeness heuristic.

The Taxicab Problem is this: a hit-and-run accident occurs at night. There are two cab companies in the city operating in distinct green and blue. 85% of cabs in the city are green cabs; 15% are blue cabs. An eyewitness identifies the cab in the hit-and-run accident as **blue**. The witness is tested under similar conditions and correctly identifies a cab's color 80% of the time, regardless of the cab's color. What is the probability that the cab involved in the accident was blue?

Kahneman and Tversky report that for several hundred subjects given slight variations of this question, both the modal and the median response to the question—"what is the the probability that the cab was blue?"—was 80%. This is almost double the correct answer of 41%. The common error is that most people focus on the eyewitness accuracy of 80%, concluding that the cab was blue with 80% probability. This neglects the significant factor of the base rate of only 15% blue cabs in the city.

The correct answer, solved using Bayes' Theorem (Equation A-3), is

[20] confirmation bias—arguably one of the most prevalent and persistent cognitive biases in human affairs—is widespread and consistently replicated in studies; it has a direct implication in decision-making under uncertainty: once an individual makes a decision, there is a strong and persistent tendency to look for confirmatory information that supports the decision

$$P(B|D) = \frac{P(D|B)P(B)}{P(B)P(D|B) + P(G)P(D|G)} = \frac{(0.8)(0.15)}{(0.15)(0.8) + (0.85)(0.2)} = 41\% \tag{10}$$

where B is the hypothesis that the cab was blue, G is the hypothesis that the cab was green, and D is the data of the eyewitness report. Given the complexity of applying Bayes' Theorem in this form, it is no wonder that most people in the experiment fail to calculate the correct answer.

Let's compare Equation 10 to the solution using Bayes Factors. The prior odds of a blue taxicab, $\Theta_B^{(0)}$, is 15/85. The Bayes Factor is given directly in the problem by the tested eyewitness reliability of 80%. That is,

$$BF = \frac{P(D|B)}{P(D|G)} = \frac{80}{20} = 4 \tag{11}$$

and thus the posterior odds of a Blue cab are

$$\Theta_B^{(1)} = BF \cdot \Theta_B^{(0)} = (4)\left(\frac{15}{85}\right) = \frac{12}{17} \rightarrow p_B = 41\% \tag{12}$$

Equation 12 is obviously more straightforward to apply than Equation 10 in solving the Taxicab Problem. This is not unusual. The direct use of Bayes' Theorem is often challenging due to the need to assess the marginal probability of observing a piece of data. In practice, this requires a self-consistent assessment of the data, the hypothesis, and the complementary hypothesis. Because the Bayes Factor is expressed as a ratio, it sidesteps the need to assess the marginal probability of the data. Furthermore, it is not sensitive to underlying assumptions or biases about the data, since these will cancel out when forming the likelihood ratio. This example offers two lessons learned on the use of Bayes Factors compared to a direct use of Bayes' Theorem.

Lesson Learned 3: BFs are often simpler to apply than direct application of Bayes' Theorem (avoiding cognitive pitfalls with conditional probabilities, *e.g.*, base rate neglect)
Lesson Learned 4: BFs produce the "right" answer without needing to assess the problematic marginal probability observed data (BFs are insensitive to assumptions about the data or evidence)

C. Is Charlie a Cheater?

This example is drawn from the author's experience while serving as Chair of the Engineering Division at the U.S. Air Force Academy. A cadet suspected that two of her classmates had cheated on a final exam, having heard them comparing answers to questions during the exam.[21] We will call our three cadets *Alice*, *Bob*, and *Charlie*. Alice was sitting in front of Bob and Charlie during a final exam. She reported to the instructor that Bob and Charlie had been talking during

[21] The Academy takes its Honor Code quite seriously; lying, stealing, or cheating may be grounds for dismissal

Figure 6: *Common Final Exam at the United States Air Force Academy*

the exam and that it sounded like they were comparing answers. Bob and Charlie did admit to talking, but denied cheating. They claimed to have been trying to coordinate a ride to the airport after the exam. So the judgment under uncertainty the author had to make: is Charlie a cheater?

Bob would eventually admit to having cheated off Charlie, but claimed that Charlie had no knowledge of Bob's cheating. Interestingly, Bob received a better grade, missing only 8 questions, compared to Charlie's 13 incorrect answers on the 50-question exam. However, of the 8 questions they jointly missed, their incorrect answers were all identical. Bob admitted to cheating and started honor probation immediately. Charlie had a previous honor violation. A second honor conviction would likely result in his dismissal, so he clearly had an incentive to deny cheating. Ultimately, Charlie was found not guilty at an honor court of his peers. He appealed his academic penalty of zero points on the exam, based on the not-guilty finding by the Honor Board. Charlie's appeal came to the author as the Chair of the Engineering Division for adjudication and final decision.

The relevant data in the case includes both exonerating and incriminating evidence. This includes:

1. Alice observed Bob and Charlie talking
2. Bob and Charlie admitted to talking, but claimed to be arranging a ride to the airport
3. Charlie was found "Not Guilty" by an Honor Board of his peers
4. Charlie's prior honor conviction provides a clear incentive to lie about cheating
5. Bob admitted to cheating off Charlie, so there is a *prima facie* case that cheating did occur
6. No other cadets sitting around Alice, Bob, or Charlie reported anything suspicious; under the Academy's Honor Code, cadets may not tolerate cheating and are obligated to confront suspicions of violations (the other cadets sitting around Bob and Charlie were also friends of Bob and Charlie)
7. There were two versions of the exams, A-versions and B-versions, alternating seat-by-seat. Bob and Charlie, sitting in adjacent seats, should have had different versions, but someone had switched the order before the exam started so that both Bob and Charlie had the same version of the exam.

8. The similarity and correlation of incorrect answers between Bob and Charlie's exams can be compared with all 186 cadets taking the same exam version

We start with two complementary hypotheses: H_0, Charlie is innocent; H_1, Charlie cheated. The Bayes Factors will be the relative likelihood of observing the data given above under these two hypotheses:

$$BF_i = \frac{P(D_i|H_1)}{P(D_i|H_0)} \tag{13}$$

Table 2 summarizes the evidence as well as the BFs the author used in judging Charlie's likelihood of having cheated. BF_1 in case of Alice's report, assuming 80% eyewitness reliability, is 4:1, just as it was in the prior taxicab example (Section IV.B). The author gave Bob and Charlie the benefit of the doubt regarding their claim that they had only been arranging a ride to the airport. Assessing equal probability of observing the talking under either hypothesis, BF_2 = 1:1.

Table 2: Is Charlie a Cheater?

	Evidence	BF_i	Comment
1	observed by student talking	4/1	≈ 80% eyewitness reliability
2	admitted to talking	1/1	gives Bob & Charlie the benefit of the doubt
3	"not guilty" at honor board	1/3	cadet & faculty confidence in the honor system
4	prior honor violation	4/2	slight incentive to lie; some benefit of the doubt
5	Bob's admission of cheating	3/2	acknowledgment that cheating did occur
6	no reports from other cadets	4/5	surrounded mostly by friends
7	duplicate adjacent exams	5/3	exam versions were deliberately exchanged
8	exam similarity	958/3	$P(x_8\|H_1) = 1$ $P(x_8\|H_0) = 3/958$ (only 3/958 students with perfect incorrect-answer correlation)

The "not guilty" verdict from the Honor Board is not as exonerating as it might otherwise be, given a persistent low confidence in the honor system among cadets at the time. The low confidence among cadets and the open admission by some cadets that they would never convict a fellow cadet make this line of evidence less compelling. Nevertheless, this was the basis for Charlie's appeal of his academic penalty, and it deserved careful consideration. The author ultimately assessed the probability that Charlie was innocent given the "not guilty" verdict as 75%, yielding a BF_3 = 1:3.

Charlie's previous honor violation and clear incentive to deny cheating is incriminating, but the author again gave Charlie a slight benefit of the doubt and assigned a relatively weak BF_4 = 4:3 to this data. Bob's admission to cheating off Charlie, allegedly without Charlie's knowledge, was dubious (particularly given the fact that Bob performed significantly better on the exam than Charlie). Still, the author also assigned a relatively weak BF_5 = 3:2 to this data. The absence of reports from other cadets in the area, all of whom were friends with Bob and Charlie, was not compelling and received a weak, exonerating BF_6 = 4:5. The deliberate exchange of alternate versions of the exam was incriminating, though Bob and Charlie denied any part in the exchange.

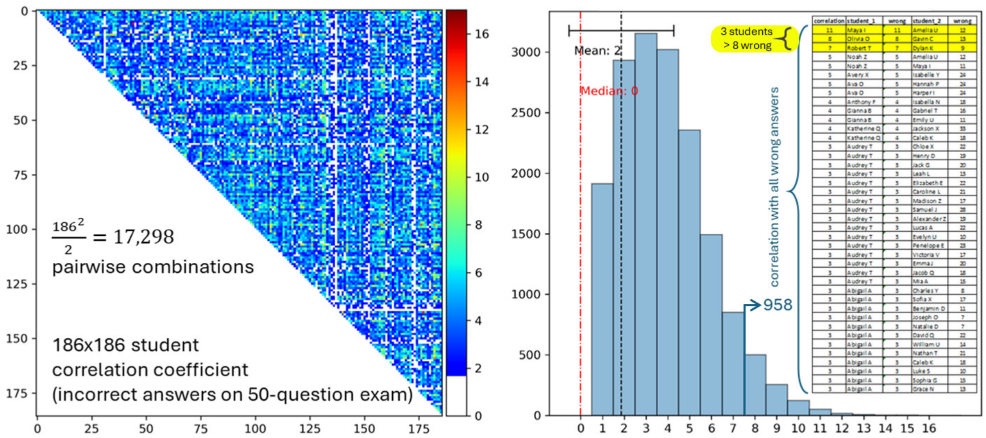


Figure 7: *(L) 186x186 pairwise correlation coefficient between all 186 students who took the final exam; (R) histogram of the 17,298 pairwise combinations between all student exams*

The author judged this to be mildly incriminating and assigned a $BF_7 = 5{:}3$.

All of the previously considered BFs in this example have been qualitative judgments given in quantitative ratios. In the case of the data on the similarity between Bob and Charlie's exams, we can be precise and quantitative. A total of 186 students took the final exam. Figure 7 depicts a heatmap of the 186x186 pairwise dot products of incorrect answers of every copy of the same version of the exam. This measure of the correlation coefficient between exams shows how similar students' incorrect answers were. A histogram of this data shows that 958 pairs of exams were correlated with eight or more incorrect answers (there were 50 questions on the exam). In only three cases—of 958 pairs of exams with 8 or more incorrect answers—did every single incorrect answer correlate exactly with the other exam. Bob and Charlie were one of these three pairs of perfectly correlated incorrect answers. The probability of observing this perfect correlation of incorrect answers, under the assumption that Charlie is a cheater, is 1. The probability of observing this data under the assumption that Charlie is innocent is 3/958 given the class-wide data. Therefore, the author assessed a Bayes Factor of 958/3 for this piece of evidence.

Taking a conservative prevalence of cadet cheating of 1%, gives prior odds of a cadet cheating of 1:99. The combined Bayes Factor is

$$BF = \left(\frac{4}{1}\right)\left(\frac{1}{1}\right)\left(\frac{1}{3}\right)\left(\frac{4}{3}\right)\left(\frac{3}{2}\right)\left(\frac{4}{5}\right)\left(\frac{5}{3}\right)\left(\frac{958}{3}\right) = 1135 \tag{14}$$

and the posterior probability that Charlie cheated in the final exam is thus

$$\Theta^{(1)} = BF \cdot \Theta^{(0)} = \left(\frac{1135}{1}\right)\left(\frac{1}{99}\right) = 11.5 \rightarrow 92\% \text{ Charlie cheated} \tag{15}$$

On the basis of this Bayesian reasoning, the author denied Charlie's request and sustained the academic penalty he had received on the Final Exam.

This example illustrates two advantages of using Bayes Factors. First, the assigned BFs indicate how strong a particular piece of evidence is judged to be. Without the exam similarity, we would generously assess the probability that Charlie cheated as less than 5%. However, the extremely strong BF_8 of more than 300 indicates how incriminating this piece of evidence is compared to the "not guilty" verdict from the Honor Board.[22] Secondly, this example shows the opportunity for constructively communicating judgment. The author was able to acknowledge and incorporate all the exonerating evidence Charlie presented. The author was also able to explain his reasoning to Charlie, along with the relative weight of the various pieces of evidence. Charlie accepted the rationality of the author's conclusions and judgments with dispassionate equanimity. As we will see in the examples of safety and technical risk discussions below, Bayes Factors provide a common framework for a meaningful discussion of the relative strengths of conflicting data, and give us a systematic way to communicate our judgments and conclusions.

Lesson Learned 5: BFs can mix both subjective and objective judgments
Lesson Learned 6: The BF-framework decomposes complex data sets into separate chunks and reveals the relative strength of different pieces of evidence (combining multiple pieces of evidence is as simple as multiplying BFs)
Lesson Learned 7: BFs provide a systematic structure for meaningful conversations and a common lexicon for communicating our experience, judgment, and decisions as we work towards consensus

D. Monty Hall Problem

This is another classic probability question [45]. The problem had been around for a while [46], but went viral in 1990, when Marilyn vos Savant shared it in her Parade Magazine column [47]. The column received thousands of letters from readers who disagreed with her answer. Many of the letters were from "experts" with PhDs; some of them were quite sexist.[23]

The Monty Hall problem is named for the host of the game show, *Let's Make a Deal* (Figure 8), in which the contestant is asked to choose between one of three doors: Door 1, Door 2, or Door 3. Behind one of the doors is a car. There are goats behind the other two. Suppose a contestant selects Door 1. The host, who knows what is behind each door, opens Door 3, revealing a goat. The host then gives the contestant the opportunity to change their choice of doors to Door 2. The somewhat counterintuitive solution, and the one that raised the ire of all the letter writers, is that switching doors doubles the contestant's chance of winning the car. The contestant, who originally had a 1/3 chance of being right with Door 1, has a 2/3 chance of winning by switching to Door 2.

Let H_i be the hypothesis that the car is behind Door i. We will refer to the data of the goat behind Door 3 as G_3. The prior odds of a car behind Door i are $\Theta_i^{(0)}$. Thus, $\Theta_2^{(0)} = 1 : 2$. The Bayes Factor for switching doors is

[22] relying on the exam similarity alone and concluding that the probability Bob and Charlie had not cheated was 3/958 would be erroneous, an example of the prosecutor's fallacy (see Section V); with the same prior odds of 1:99, the evidence from the exam similarity alone only gives 3:1 posterior odds that Charlie cheated
[23] she published many of the letters in a subsequent column, with only a trace of gloating [48]

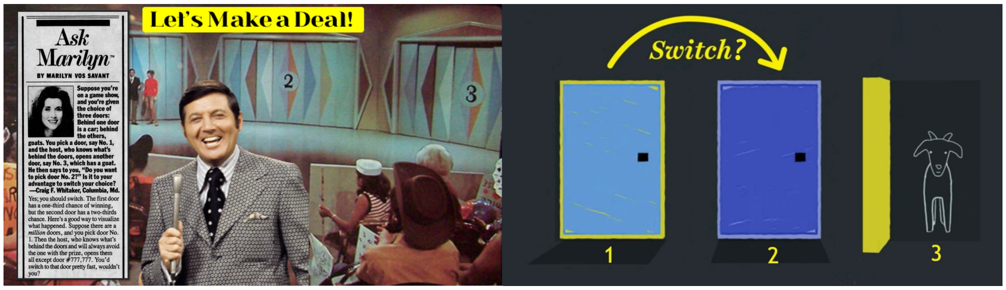


Figure 8: *The Monty Hall Problem. Do you switch doors?*

$$BF = \frac{P(G_3|H_2)}{P(G_3|H_2^c)} = \frac{P(G_3|H_2)}{\frac{1}{2}P(G_3|H_1) + \frac{1}{2}P(G_3|H_3)} = \frac{1}{\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)(0)} = \frac{1}{1/4} = 4 \tag{16}$$

since the probability of observing G_3 under the assumption of H_2 being true is 1 given the problem set-up (if Monty reveals a goat to the contestant given H_2, the only choice is G_3). The probability of observing that data under the assumption of H_2^c, or $\neg H_2$, is one out of four (Figure 9). This is because the hypothesis H_2^c must consider both H_1 and H_3, which may be considered equally probable. The posterior odds for switching doors are thus

$$\Theta_2^{(1)} = BF \cdot \Theta_2^{(0)} = \left(\frac{4}{1}\right)\left(\frac{1}{2}\right) = 2 \rightarrow 66\% \text{ winning (switching to Door 2)} \tag{17}$$

There is an alternative, more direct BF method of obtaining the correct answer. Consider a comparison of the odds between Door 1 and Door 2. The prior probability for either door is 1/3, so the prior odds ratio for Door 1 to Door 2 is 1:1. The Bayes Factor comparing the likelihood of Door 1 to Door 2 given the goat behind Door 3 is

$$BF_{D_1:D_2} = \frac{P(G_3|H_1)}{P(G_3|H_2)} = \frac{1/2}{1} = \frac{1}{2} \tag{18}$$

thus, the posterior odds for Door 1 to Door 2 are

$$\Theta_{D_1:D_2}^{(1)} = BF_{D_1:D_2} \cdot \Theta_{D_1:D_2}^{(0)} = \left(\frac{1}{2}\right)\left(\frac{1}{1}\right) = \frac{1}{2} \rightarrow 33\% \text{ winning (keeping Door 1)} \tag{19}$$

To make things more interesting, and reflecting the definition that *probability* is our subjective relationship with uncertainty, suppose the contestant has additional data, considerations, or suspicions about Monty's behavior. *E.g.*, consider the case that, through repeated observation of *Let's Make a Deal*, the contestant concludes the car is more often behind Door 1. Or perhaps the contestant suspects that the producers do a poor job of randomizing doors and never repeat the same door for the car from the previous week's episode. Or perhaps we suspect that Monty knows about the solution to the Monty Hall problem, will only offer the contestant the chance to switch if

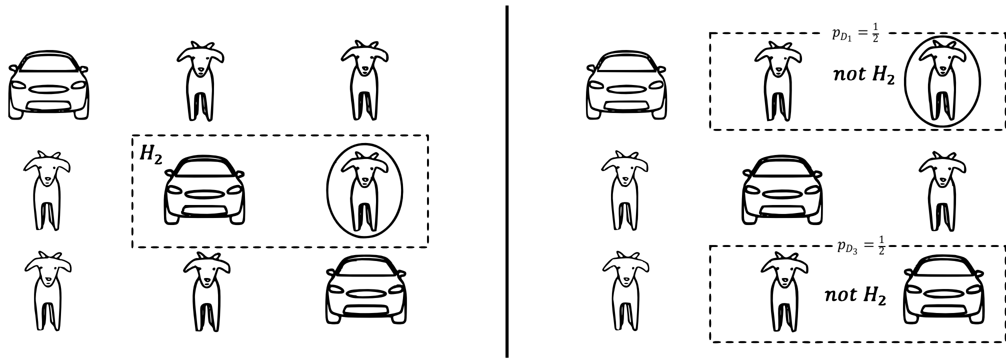


Figure 9: *(L) H_2: hypothesis that the car is behind Door 2;*
(R) not H_2: hypothesis that the car is not behind Door 2

they initially selected the correct door, or will double down and offer the contestant cash instead. These are all subjective judgments or observation-informed expectations about the real world that the contestant can systematically quantify using Bayes Factors.

Lesson Learned 8: BFs can account for and incorporate other factors that were not part of the original problem statement

E. Project Serene

Project Serene was a non-traditional, long-range weapon test with considerable unknowns and uncertainty. The weapon had a limited pedigree of previous employment, but the platform and method of employment under test were entirely new. For various programmatic reasons, many of the typical risk reduction and build-up tests were not possible for the project. These were still listed as potential evidence, indicating how the posterior odds could be influenced by additional data, but were not assigned BFs in computing the posterior odds. The initial data and initial planning BFs are given in Table 3.

The detailed rationale for each BF in Table 3 is not as important as the technique. Figure 10 contains the author's notes from the initial planning meeting. This demonstrates how straightforward it is to track multiple lines of evidence during a meeting. Following extended discussion regarding each factor with all the experts at the table, the author recorded his real-time assessment of the Bayes Factor, which was the likelihood of that factor under the hypothesis that the test would be safe relative to that factor under the hypothesis that the test would be unsafe. Overall, the approach was a straightforward "piece of scratch paper" technique and revealed the relative strength of different factors under consideration.

The overall initial Bayes Factor from the initial planning meeting for the factors in Table 3 was 32. With a "generous" prior odds ratio of 1:100, the initial posterior odds are

$$\Theta^{(1)} = BF \cdot \Theta^{(0)} = \left(\frac{32}{1}\right)\left(\frac{1}{100}\right) \approx \frac{1}{3} \rightarrow 25\% \text{ (expectation of safe test)} \tag{20}$$

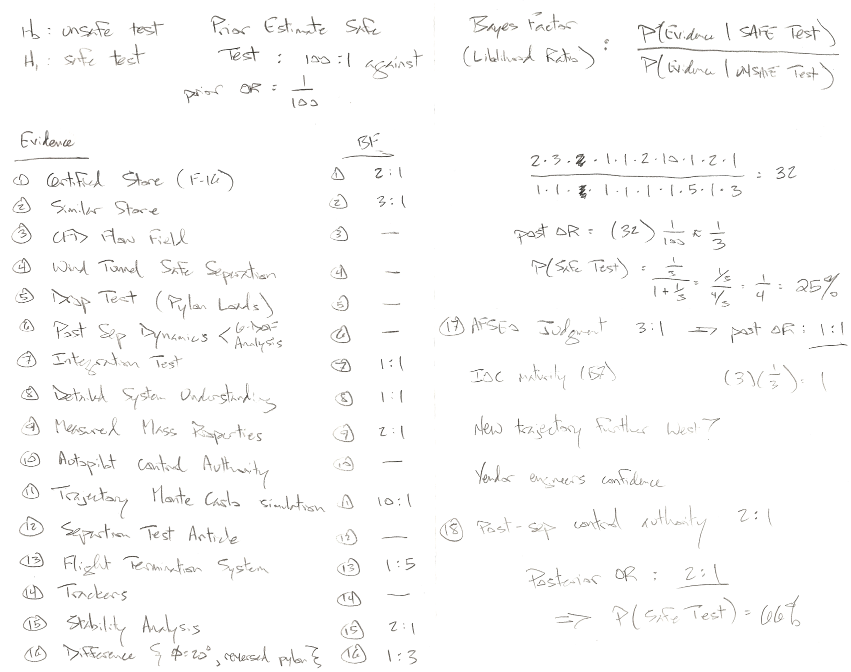


Figure 10: *Project Serene TAB/SRB notes showing "back of the envelope" Bayes Factor calculations*

Table 3: Project Serene - Initial Bayes Factors

	Evidence	BF	Comment
1	certified store	4:1	but different platform
2	similar store	3:1	engineering data supports similarities
3	CFD flow field predictions	—	not accomplished
4	wind tunnel safe separation test	—	not accomplished
5	drop test (pylon loads)	—	not accomplished
6	post-sep dynamics (6-DOF analysis)	—	not accomplished
7	integration testing	1:1	limited scope
8	detailed system understanding	1:1	lack of engineering data & technical detail
9	measured mass properties	2:1	somewhat limited
10	'autopilot' control authority	—	unknown
11	Monte Carlo trajectory analysis	10:1	acceptable; unfavorable edge cases indicate potential off-range impact
12	separation test article	—	not possible
13	flight termination system	1:5	not feasible
14	trackers	—	not feasible
15	stability analysis	2:1	statically stable article
16	reversed pylon installation	1:3	uncertainty from differences

During the planning meetings, the author also routinely solicited BFs from the other meeting participants. A significant range of opinions emerged regarding the safety of the proposed test. The lack of expert consensus provided a direct measure of the overall uncertainty, underscoring a wide confidence interval for possible outcomes. This uncertainty measure is discussed in more detail below. Having meeting participants express their odds ratios and Bayes Factors also revealed whether the difference in opinion was due to the strength (or risk) of a particular factor or due to a difference in their prior odds of a safe test.

The author's initial assessment of a posterior odds of 1:3, or 25% probability of a safe test, prompted a request for additional information. In particular, the author, serving as the Test Acceptance Authority for Project Serene, requested expert opinions and judgments from the Air Force SEEK EAGLE Office (AFSEO). AFSEO provided further analysis, along with their judgment, providing additional data and prompting an updated Bayes Factor (Table 4).

Table 4: Project Serene - Bayes Factors II

	Evidence	BF	Comment
17	AFSEO judgment	3:1	lack of engineering data results in significant residual uncertainty; store similarity alleviates some risk based on engineering judgment

The updated posterior odds given AFSEO's recommendation are given by the product of the previous posterior odds (the 'new prior') and the new Bayes Factor:

$$\Theta^{(2)} = BF_2 \cdot \Theta^{(1)} = \left(\frac{3}{1}\right)\left(\frac{1}{3}\right) \approx 1 \rightarrow 50\% \text{ (expectation of safe test)} \tag{21}$$

Given even posterior odds following AFSEO's updated analysis, the author was still unwilling to approve the test. The most significant residual concerns were due to uncertainty over the post-separation control authority of the system under test and the kinematic potential for the weapon to depart the range and impact a populated area. This prompted a request for additional engineering data from the vendor and discussions with the engineering team. The additional evidence resulted in a third Bayes Factor (Table 4).

Table 5: Project Serene - Bayes Factors III

	Evidence	BF	Comment
18	post-separation control authority	2:1	some confidence in post-separation dynamics & updated kinematic analysis

The updated posterior odds considering the additional data are given by the product of the previous posterior odds and the new Bayes Factor:

$$\Theta^{(3)} = BF_3 \cdot \Theta^{(1)} = \left(\frac{2}{1}\right)\left(\frac{1}{1}\right) \approx \frac{2}{1} \rightarrow 66\% \text{ (expectation of successful test)} \tag{22}$$

The shift in assessment from "safe" to "successful" was unavoidable. Safety risk and technical risk are often troublesome to separate as they are generally not mutually exclusive. With Project Serene, the substantial residual uncertainty of the system under test meant that considerable accompanying uncertainty existed about the test's technical success as well.

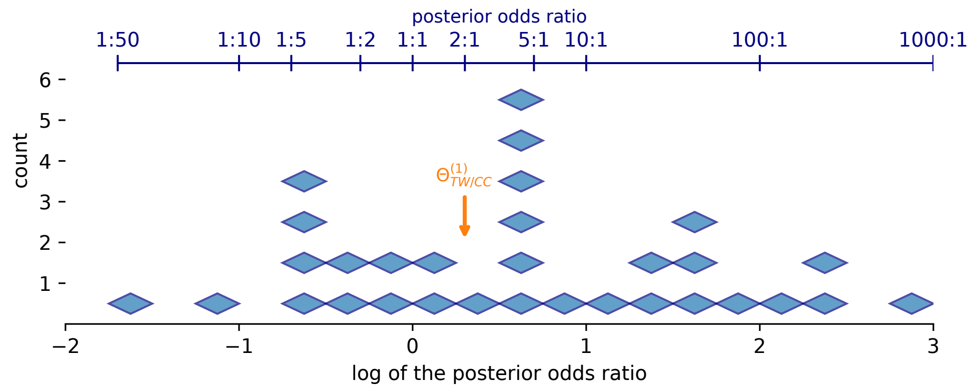


Figure 11: *Project Serene posterior odds ratios for all 30 safety review participants*

At this point, after several months of planning, the author asked all participants (approximately thirty aircrew, test safety, range safety, test engineers, project planners, program engineers, and technical advisors) to give their own final, individual posterior odds ratio that the test would be successful. The results, Figure 11, indicate almost five orders of magnitude in differing opinions. The author was ultimately *more-confident-than-not* that the test would be safe and approved the test (the author's ≈2/3 confidence is indicated in Figure 11).

Ultimately, the first launch was not successful due to a technical deficiency, although the overall test was not unsafe. The second attempt following additional work was successful. A single example is not definitive, but the author finds the consistency between the Bayesian reasoning in this example, the range of test planning expectations, and the eventual outcome to be consistent and representative.

> **Lesson Learned 9** : BFs are straightforward to perform real-time and via "back-of-the-envelope" calculations
> **Lesson Learned 10**:BFs provide a common framework to express and expose the full range of subject matter expert opinions; the range of opinions provides a good estimate for uncertainty and for bounding judgments with confidence intervals

F. B-21 First Flight

During the First Flight Readiness Review for the B-21 in 2023, the author used Bayes Factors to assess technical, programmatic, and security risks in addition to safety risks. Although we will not provide details for security reasons, the application was straightforward and incorporated data from ground testing, taxi testing, aircrew and test team training, mission rehearsals, airworthiness, and engineering reviews. Sequential updates to the posterior odds were made over the course of several weeks, providing a useful framework for communicating judgments and opinions.

> **Lesson Learned 11**: BFs provide a framework to expose a range of opinions, which in turn can provide a basis for uncertainty and forming confidence intervals of our ultimate judgments

Figure 12: *B-21 First Flight in the fall of 2023*

G. Hermeus Quarterhorse Mk1

Hermeus is a relatively new aerospace and defense technology company backed primarily by venture capital. Hermeus' initial efforts, the Quarterhorse program, are focused on incrementally developing and building high-Mach and hypersonic aircraft capable of reaching Mach 5+ speeds. The Hermeus Quarterhorse Mk1 is an uncrewed, remotely piloted aircraft powered by a single GE J85 turbojet engine. In May 2025, the Quarterhorse Mk1 had a successful first flight at Edwards AFB after a year of rapid development and fast-paced ground testing at Edwards.

Table 6: Hermeus Quarterhorse Mk1 initial planning meeting Bayes Factors

	Evidence	BF	Comment
1	taxi testing (AEDC)	3:1	accomplished
2	taxi testing (Edwards)	—	to be accomplished
3	comm-link range testing	—	to be accomplished
4	wind-tunnel stability data	5:1	good results from USAFA's subsonic wind tunnel
5	lack of autopilot stabilizing	1:4	unknown initial trim settings
6	FTS	1:1	hybrid design, need additional info
7	glide-cone footprint	10:1	kinematic constraints are strong risk mitigation
8	lakebed landing option	5:1	lack of engineering data
9	datalink transfer to alt mode	1:3	somewhat limited
10	engine runs	—	need x-wind data for limits
11	Airworthiness	—	pursuing alternate pathway with FAA

When Hermeus' leadership first approached the 412th Test Wing (TW) to discuss testing the Quarterhorse Mk1 in Edwards AFB's R-2515 restricted airspace, they shared a willingness and tolerance to accept risk as a means to test and learn quickly. The Mk1 program was assigned to the 412 TW's Experimental Test Force (ETF) and given the mandate to support the Hermeus First

Figure 13: *Hermeus Quarterhorse Mk1 undergoing taxi testing at Edwards AFB*

Flight test campaign. Bayes Factors proved to be a useful method for continuously assessing and communicating the author's opinion of the readiness of the Mk1 for First Flight. Table 6 includes the Bayes Factors the author used at the very first, executive kickoff meeting with Hermeus leadership. Placeholders for Bayes Factors for testing that would be accomplished when the Mk1 arrived at Edwards, *e.g.*, taxi testing and command-and-control link range testing. This served as useful "good-faith" confidence during that initial planning meeting.

Table 7: Hermeus Quarterhorse Mk1 Test Approval Board

	Evidence	BF	Comment
1	Team & Engineering	10:1	impressive, small team with trusted pedigreee and demonstrated experience & competence
2	successful taxi testing	4:1	to be accomplished
3	datalink margins	2:1	to be accomplished
4	design & certified parts	1.5:1	
5	lack of stab augmentation	1:1.5	design uncertainty; unknown initial trim settings
6	footprint containment	10:1	kinematically contained to remain within the boundaries of the lakebed
7	FTS	5:1	independent, but non-redundant fuel-cutoff system
8	loss link logic	5:1	lack of engineering data
9	Airworthiness	—	pursuing Title 14 exemption

Following taxi testing in the fall of 2024, the initial Bayes Factor for the evidence considered at the Test Approval Board (TAB) is given in Table 7. The BF of 20,000 combined with the a prior odds ratio of 10^{-3} is

$$\Theta^{(1)} = BF \cdot \Theta^{(0)} = (20,000)\left(\frac{1}{1000}\right) = 20 \rightarrow 95\% \text{ (expectation of safe test)} \tag{23}$$

The author's risk assessment was based primarily on a responsibility to protect government infrastructure and other test assets. Although the author respected and supported the Hermeus Team's tolerance of elevated risk, the author could not accept risk to other high-priority programs, *e.g.*, B-21 testing. (To put it crudely: it was ok if Mk1 crashed on its first flight, but it was not acceptable for it to crash into the B-21 compound.) Bayes Factors proved useful for articulating and quantifying the most significant factors, providing a transparent basis for meaningful, technical discussions between team members and the test approval authority.

The completion of taxi testing and test approval coincided with the arrival of winter precipitation at Edwards AFB and the closure of the lakebeds. One of the risk mitigations for approval was a restriction to lakebed-contained operations. The closure of the lakebeds resulted in an extended program delay. During the three-month delay between test approval and the ultimate first flight, several key test personnel left the Mk1 program, driving a follow-on first-flight readiness review. The factors considered are given in Table 8.

Table 8: Hermeus Quarterhorse Mk1 FINAL First Flight Review

	Evidence	BF	Comment
10	long lay off	1:3	3-mo break due to red lakebeds
11	personnel turnover	1:4	departure of key Hermeus personnel (reduction to previous strong credit)
12	mitigation	2:1	observed/reviewed test conduct to build confidence
13	FAA airworthiness waiver	1:1	waiver given; no analysis implied
14	team currency	2:1	mission rehearsals, training, EP sims
15	company/programmatic pressure	1:2	unavoidable; good potential for "Drift"

The preponderance of factors considered during the first flight readiness review indicated an increased risk relative to the earlier assessment. Using the prior odds from previous test acceptance posterior of 20:1 and the BF_{FFRR} from Table 8 of 1/6th, the final posterior odds are

$$\Theta^{(2)} = BF_{FFRR} \cdot \Theta^{(1)} = \left(\frac{1}{6}\right)\left(\frac{20}{1}\right) = \frac{10}{3} \rightarrow 77\% \text{ (expectation of safe test)} \tag{24}$$

As with Project Serene (Section IV.E), the iterative use of Bayes Factors was useful in continuously updating our judgment based on new information. This is the essence of '*Being Bayesian*.' The systematic record of Bayes Factors from earlier meetings was also helpful in recalling the basis of our previous judgments (*e.g.*, appropriately discounting the earlier credit given for trust in key personnel following their departure from the current test effort) and for quantifying and communicating our judgments.

Lesson Learned 12: BFs are useful for tracking changing judgments over time and for communicating those judgments with stakeholders
Lesson Learned 13: BFs are useful for communicating status and progress towards an objective as well as the relative importance of factors for "getting to yes" (avoiding a sequence of "bring me a rock" drills)

H. C-17 Missile Defense Agency High-Altitude Airdrop

The air-launched intermediate-range ballistic missile (AL-IRBM) is a target missile developed by the Missile Defense Agency (MDA) to test the Ballistic Missile Defense Shield (BMDS). The AL-IRBM is airdropped and launched from a C-17A to simulate an Intercontinental Ballistic Missile (ICBM) trajectory. After launch, the AL-IRBM is targeted and intercepted using the missile defense shield. The high-altitude (25k ft) airdrop from the C-17 requires deliberate aircraft depressurization, exposing crew members to physiological risks of hypoxia and decompression sickness ('the bends'). The danger is compounded by the remote operational areas required for the mission, with the aircraft typically operating more than five hours away from medical facilities equipped with hyperbaric chambers. Delaying treatment for decompression sickness (DCS) beyond four hours significantly increases the risk of catastrophic adverse outcomes for DCS.

In 2023, a 33-year-old civilian employee of the Missile Defense Agency died of cardiac arrest after experiencing decompression sickness during an Air Mobility Command (AMC) C-17 high-altitude airdrop mission in Alaska. Following this Class A mishap, responsibility for all future high-altitude C-17 MDA missions was returned to Air Force Test Center and the 418 FLTS was assigned the Participating Test Organization responsibility.

During regular review and approval of the Test and Safety planning package for MDA missions in 2025, the risk of DCS, one of four long-standing test hazards in the safety package, was reassessed as HIGH RISK due to new data from the Air Force's 711 Human Performance Wing (HPW). Given almost two decades of experience with similar airdrop missions, the majority of which had previously been designated MEDIUM RISK, there was substantial disagreement among subject matter experts on the Safety Review Board (SRB) regarding the appropriate level of risk. An approach using Bayes Factors proved to be a useful framework for building a common lexicon, understanding the various perspectives, and ultimately reconciling the recommendation.

The SRB considered two different sets of data during the safety assessment: previous MDA test program flight physiological data and DCS research from the 711 HPW and the dive community. From 24 previous MDA test campaigns since 2004, with an average of 4-5 flights per program and approximately 20 personnel per flight, the SRB estimated there were 2000 previous person-flight hours of high-altitude exposure (approximately one hour per person per mission). During these 2000 previous person-flight hours, there were seven reports of physiological incidents: six involving hypoxia or hyperventilation and the MDA civilian fatality the day after the mission while being treated for DCS.

The USAF's Altitude DCS Risk Assessment Computer (ADRAC) predicts a 1% chance of DCS for the test altitude and exposure time.[24] The SRB's independent reviewer from Aerospace

[24] 22,000 ft standard day for 60 min

Figure 14: *418 FLTS C-17 taking off from Edwards AFB & deploying an air-launched medium-range ballistic missile*

Flight Medicine agreed with ADRAC's estimate of a 1% chance of DCS and added the expectation of a 1% chance of a catastrophic outcome for any given case of DCS based on a comprehensive review of data from the diving community. The probability of a catastrophic event is then estimated to be 1% of 1%, corresponding to a probability of death or severe injury of 10^{-4}. This is within the 90% confidence interval (9×10^{-5} to 2×10^{-3}) for the 'actual' probability based on the observed rate of 1 catastrophic event per 2000 person-flight hours (5×10^{-4}).[25]

Based on the number of personnel flying and the number of high-altitude, unpressurized flights in the test program, the SRB concluded an overall probability of a catastrophic event during the program of 1/250 (0.4%).[26] This corresponds to a mishap severity-probability level of ***unlikely/catastrophic*** (1/C - HIGH RISK) per AFTCI 91-202 [10]. Based on these probability calculations, most SRB members felt constrained by the clear-cut numerical statistics from the ADRAC calculations. Other SRB members, particularly those with extensive experience flying operational high-altitude missions (*e.g.*, aircrew and loadmaster), disputed a HIGH RISK characterization on the basis that "high-altitude airdrop is a routine operational mission." Table 9 summarizes the various opinions of the SRB.

Table 9: C-17 high-altitude airdrop SRB judgments

Risk	severity level	mishap probability level	Comment
1/C	*catastrophic*	*unlikely*: 1-in-1000 to 1-in-100	$p = 1/250$ estimate
1/D	*catastrophic*	*highly unlikely*: 1-in-10^6 to 1-in-1000	$\Theta^{(0)} < 10^{-3}$
2/C	*critical*	*unlikley* : 1-in-1000 to 1-in-100	$BF > 10$

From a Bayesian perspective, it is easy to understand and separate the viewpoints into three categories: those who agreed with the ***unlikely/catastrophic*** (1/C - HIGH RISK) categorization; those who doubted whether death was truly credible given the mitigations in place and favored an ***unlikley/critical*** (2/C - MED RISK) categorization; and those who, although acknowledg-

[25] we use a Gamma-Poisson conjugate prior (see Appendix B) to create a Bayesian estimate for the distribution of the rate of a rare event. With the Jeffreys (uninformed) prior $\sim \Gamma(0.5, 0)$ for the rate parameter λ, the posterior distribution becomes $\Gamma(1.5, 2000)$ after observing a single event in 2000 exposures (Figure B-3)

[26] 4 flights with 10 personnel per flight yields 40 total exposures; the probability of at least one event in 40 trials is $1 - (1 - 10^{-4})^{40} = 0.004$

ing death as a credible consequence based on prior MDA mission experience, could not justify a subjective probability greater than 10^{-3} (*highly unlikey*) and favored maintaining the prior risk assessment of ***catastrophic/highly unlikely*** (1/D - MED RISK).

Differences of opinion or judgment among subject matter experts are not uncommon. Discounting deliberate 'risk-hacking' to 'back-into' a pre-determined risk category, the views and opinions of experts are usually sincere and genuinely held. Constructive conversations that achieve a consensus can be difficult or impossible if we lack a common framework or lexicon. Bayes Factors offer a solution. They provide a common lexicon by which to have meaningful discussions. Under this mental model for decision-making, differences of opinion are attributed to either differences in priors or differences in likelihood ratios. With the MDA test, those favoring 1/D categorization had a different prior. Those who favored a 2/C categorization had a BF that strongly weighed the mitigations. Those who felt "bound by" the 1/C categorization based on the clear-cut mathematics were offered an opportunity for additional agency with a Bayesian approach that let members update their odds based on data.

We do not have to reach consensus in flight test; we have ultimate risk acceptance authorities whom we ask to make the final call. However, we need to provide our teams with the means to have meaningful and constructive discussions and to hear the diverse opinions that are key to realizing the wisdom of the crowd. Bayes Factors are a good framework for providing that structure.

Lesson Learned 14: BFs offer a framework with a common lexicon for discerning underlying reasons for differences in opinion and judgment (differences in posterior judgments reflect either a different prior or a different likelihood ratio)

V. Discussion & Lessons Learned

The examples in Section IV were deliberately selected to highlight the utility and the use of Bayes Factors. Bayesian reasoning is a useful and reliable method for updating our understanding of an uncertain world; however, directly applying Bayes' rule in practical situations is often challenging due to the difficulty of estimating the probabilities of the observed data or evidence. Bayes Factors sidestep this difficulty. Additionally, systematic errors or erroneous assumptions about the underlying data are avoided because, as long as the same set of errors and assumptions is present in assessing the Bayes Factor, they cancel out. Even when the probabilities are clear, using Bayes Factors and odds ratios is frequently simpler than applying Bayes' Theorem in conditional probability form (reference the examples in Sections IV.B and IV.D).

The posterior odds/probabilities are dependent on the prior odds/probabilities. The selection of prior probabilities is a persistent and sometimes contentious aspect of Bayesian analysis. Some critics are quick to dismiss the overall approach by claiming that priors are "just made up," thereby introducing arbitrary subjectivity into supposedly objective risk assessments. ET Jaynes[27] describes the problem: "An unfortunate impression has been created that rejection of personalistic

[27] Edwin Thompson Jaynes (1922–1998): American physicist and mathematician who fundamentally transformed probability theory and statistical inference; "Probability Theory: The Logic of Science" [20], is one of the most influential works in modern Bayesian statistics

probability automatically means the rejection of Bayesian methods in general. It will hopefully be shown here that this is not the case; the problem of achieving objectivity for prior probability assignments is not one of psychology or philosophy, but one of proper definitions and mathematical techniques, which is capable of rational analysis" [49].

The criticism of subjectivity in priors fundamentally ignores the nature of probability and the reality of decision-making under uncertainty. In Bayesian reasoning, probabilities are not frequencies waiting to be discovered through repeated experiments, but expressions of our current state of knowledge and degree of belief about uncertain propositions. When a flight test team assigns a prior probability estimate for a test event, they are not arbitrarily picking numbers but systematically encoding their individual and collective engineering judgment. The opportunity to incorporate and include experience with similar systems, similar tests, analysis of previous test results, understanding of team readiness, and an understanding of the physical principles of the test in the prior offers a genuine assessment of what the team knows before conducting the test. The value in making explicit the assumptions and beliefs that would otherwise remain hidden is that they may be openly discussed by the team using a common framework and lexicon. Additionally, one of the mathematical properties of Bayesian inference ensures that with sufficient evidence, the choice of prior is irrelevant: different priors will converge to the same posterior distribution as data accumulates.

The real question is not whether we should use subjective priors, but whether we should make our inevitable subjectivity explicit and transparent. Without an explicit expression of our judgments, we leave them implicit and uninfluenced by the careful consideration of evidence. With an explicit Bayesian approach, our subjectivity is available for systematic revision. By exposing what we know or think we know, and by having a common reference point to compare our judgments with others, we gain a clear advantage from adopting a formal Bayesian approach in our flight test safety planning.

Differences of opinion should be expected and encouraged. With a Bayes Factor mental model of decision-making, the differences of opinion are either due to differences in priors or differences in likelihood ratios (BFs) based on the relative strength of the evidence. In revealing differences in prior odds or likelihood ratios, and in articulating our reasoning and judgment, we establish a framework for having meaningful and productive conversations and building a consensus that leverages the collective wisdom of the crowd.

Many of the cognitive errors and biases described in Section II are compounded when considering complex problems with complicated descriptions. Concise problem statements and decomposition of complex problems into simpler, more manageable, cognitively concise components can aid our analysis and judgment. By chunking data and evidence into separate pieces for piecewise consideration, Bayes Factors can help alleviate the overwhelming weight of trying to cognitively digest all the evidence at once for some problems. Additionally, by parsing our separate judgments and by making clear distinctions between priors and likelihoods, we can avoid cognitive traps such as the "prosecutor's fallacy."

The prosecutor's fallacy is a critical cognitive error that confuses a statement about the likelihood ratio (the Bayes Factor) with a statement about the posterior probability given the evidence. For example, the prosecutor incorrectly switches $P(D_{DNA}|H_{Innocent})$ for $P(H_{Innocent}|D_{DNA})$. The fallacious prosecutor argues that since the probability that a DNA sample found at a crime scene matches a random person is 1-in-10-million, then the probability that the defendant is innocent given a DNA sample match is 1-in-10-million. An equivalently erroneous argument would be swapping the statement "most popes are catholics" with "most catholics are popes." The prosecutor's fallacy is a form of base rate fallacy and it has led to wrongful convictions and other miscarriages of justice.[28] In explicitly expressing our hypotheses, Bayes Factors, and odds ratios as we recommend here, we clarify otherwise complex statements and are less likely to make erroneous judgments due to confusing conditional probability statements.[29]

Finally, because we are forced to consider alternative hypotheses when forming likelihood and odds ratios, we are less likely to be surprised by the unexpected event and may be less likely to fall victim to confirmation bias. BFs will not eliminate confirmation bias or other cognitive errors. But, by forcing us to be explicit about alternative hypotheses—and by assigning a numerical value to alternative outcomes—we have a better chance of avoiding the preferential tendency to see data that confirms our prior beliefs. When making decisions under uncertainty, practicing extreme humility and protecting a willingness to admit we may be wrong is a good guard against hubris. Guarding against hubris is a healthy starting point for good judgment.

Does flight test really need another new approach? Bayes Factors do not really represent a fundamentally different way of approaching risk. BFs are merely a different way of organizing, communicating, and updating our state of knowledge and uncertainty. Having a common lexicon grounded in firm mathematical foundations is useful. As such, the 412 TW Test Safety, which teaches the test safety planning at the USAF Test Pilot School, has started incorporating a Bayes Factor approach in the TPS curriculum. More work and education are necessary before Bayes Factors are widely adopted and routinely employed, but the utility for the flight test community is clear and will likely endure.

A. Collected Lessons Learned on the Practical Use of Bayes Factors BFs:

1. Our judgments and conclusions are strongly influenced by our prior expectations
2. The BF-framework requires us to acknowledge alternate hypotheses explicitly (and may guard against confirmation bias, blind spots, and other cognitive errors)
3. BFs are often simpler to apply than direct application of Bayes' Theorem (avoiding cognitive pitfalls with conditional probabilities, *e.g.*, base rate neglect)
4. BFs produce the "right" answer without needing to assess the problematic marginal probability observed data (BFs are insensitive to assumptions about the data or evidence)
5. BFs can mix both subjective and objective judgments

[28] the OJ Simpson and Sally Clark murder trials both include good, cautionary examples of the defense fallacy and prosecutor's fallacy [50, 51]

[29] another example of the prosecutor's fallacy and the use BFs to avoid an incorrect conclusion: consider a doping test which is 95% accurate, so that $P(+\,test|doper) = 0.95$ and $P(-\,test|clean) = 0.95$. This gives a $BF = \frac{0.95}{0.05} = 19$. If the prevalence (proportion) of dopers in a sport is 1/50, then the posterior probability that someone is a doper given a positive test is only $(1/49)(19) = (19/49) \approx 28\%$, not 95% (the erroneous claim via the prosecutor's fallacy)

6. The BF-framework decomposes complex data sets into separate chunks and reveals the relative strength of different pieces of evidence (combining multiple pieces of evidence is as simple as multiplying BFs)

7. BFs provide a systematic structure for meaningful conversations and a common lexicon for communicating our experience, judgment, and decisions as we work towards consensus

8. BFs can account for and incorporate other factors that were not part of the original problem statement

9. BFs are straightforward to perform real-time and via "back-of-the-envelope" calculations

10. BFs provide a common framework to express and expose the full range of subject matter expert opinions; the range of opinions provides a good estimate for uncertainty and for bounding judgments with confidence intervals

11. BFs are also applicable and useful for making technical, programmatic, schedule, and other risk assessments

12. BFs are useful for tracking changing judgments over time and for communicating those judgments with stakeholders

13. BFs are useful for communicating status and progress towards an objective as well as the relative importance of factors for "getting to yes" (avoiding a sequence of "bring me a rock" drills)

14. BFs offer a framework with a common lexicon for discerning underlying reasons for differences in opinion and judgment (differences in posterior judgments reflect either a different prior or a different likelihood ratio)

VI. Conclusion

There is no such thing as absolute certainty, but there is
assurance sufficient for the purposes of human life

John Stewart Mills

Risk management, in flight test or otherwise, is about exercising judgment and making decisions under uncertainty. All risk management frameworks and decision-making models incorporate an explicit or implicit assessment of the probability of success or failure for a particular outcome. Accurate probability assessments for many practical problems of epistemic uncertainty in the "real-world" are difficult, with even the best engineering judgments often varying by several orders of magnitude.

Guided by the twin observations that 1) probability does not exist and 2) the wisdom of the crowd is effective, we embraced Bayes Factors as a means of having a common framework and lexicon to improve our judgment of risk and decision-making under uncertainty. Bayes Factors provided a means to practice Bayesian reasoning, systematically incorporating new data and evidence, while sidestepping some of the challenges of assessing the marginal probability of the data.

In the use of Bayes Factors, we found value in being forced to state our priors, revealing our assumptions explicitly. By explicitly admitting and considering alternative hypotheses, we expect to benefit from a decreased likelihood of being surprised by the unexpected. We found clarity in discussing and acknowledging our mental models when describing likelihood ratios. Bayes Factors will not altogether eliminate cognitive biases such as confirmation bias, but they do help guard against them.

We should never forget that uncertainty will always lie at the heart of flight test. Our goal is to be intelligent and full of doubt—to make the best decisions under uncertainty that we can, and to have the extreme humility to admit that we could be wrong. Becoming Bayesian is a good guard against hubris and being cocksure. We must first not fool ourselves... and we are the easiest person to fool.[30]

Acknowledgments & Disclaimer

I have spent two decades thinking about risk management and am indebted to many in our flight test community for the discussions, ideas, feedback, critiques, and dedication to making our profession as safe and rigorous as possible. Since learning about the Bayes Factor approach while serving as a Permanent Professor at USAFA, I have significantly benefited from multiple, extended intellectual exchanges on the topic. These helped inform and clarify my understanding, practice, and explanation. Additionally, I am grateful for the patience and indulgence of the men and women of the 412th Test Wing as we practiced and implemented the application of Bayes Factors in our flight test risk management at Edwards AFB.

It is a challenge to fully credit everyone who has helped to hone the ideas in this paper. Many have contributed in many different ways. In particular, I am grateful for the discussions and feedback from Maj Gen Scott “Nova” Cain, Brig Gen (ret) Matthew Higer, Mr Art Huber, Mr Tim “Batman” West, Col Matthew “KITT” Caspers, Col (ret) Andy “Sonic” Freeborn, Col James “Nut” Gresham, Col James “T-Pain” Hayes, Col Dan “Animal” Javorsek, Col (ret) Mark “SCIPR” Jones Jr., Col Maryann “Yetti” Karlen, Col Brian “Bandit” Neff, Col Ryan “Hulk” Sanford, Col “Fangs” Valpiani, Mr Nathan CAP’N Cook, Mr Kirk Harwood, Mr Wei “FUG” Lee, Mr Chris “NORAD” Liebmann, Mr Dan “Dan-O” Osburn, Mr Dave VanHoy, Lt Col (ret) Tyler “Matrix” Robarge, Lt Col Timothy “Speed” Lau, Lt Col Dan Edelstein, Lt Col Matthew “Fear” Gray, Lt Col Robert “Bear” Newton, Lt Col Carlos “Yardman” Pinedo, and Lt Col (ret) Donald “Whiz” Sheesley.

[30] Richard Feynman’s advice to the Caltech graduating class in 1974: “The first principle is that you must not fool yourself–and you are the easiest person to fool” [52]

References

[1] Douglas P Wickert. "Risk Awareness: A New Framework for Risk Management in Flight Test". In: *Proceedings to the SETP 62nd Annual Symposium* (2018).

[2] Mica R Endsley. "Design and evaluation for situation awareness enhancement". In: *Proceedings of the Human Factors Society annual meeting*. Vol. 32. 2. Sage Publications Sage CA: Los Angeles, CA. 1988, pp. 97–101.

[3] U.S. Department of Defense. *MIL-STD-882E w/CHANGE 1, Department of Defense Standard Practice for System Safety*. Office of the Under Secretary of Defense for Acquisition and Sustainment, July 2023. URL: https://www.cto.mil/wp-content/uploads/2025/07/MIL-STD-882E-w_CHANGE-1.pdf.

[4] D.J. Spiegelhalter. *The Art of Uncertainty: How to Navigate Chance, Ignorance, Risk and Luck*. W.W. Norton & Company, 2025. ISBN: 9780241658635.

[5] Amos Tversky and Daniel Kahneman. "Judgment under Uncertainty: Heuristics and Biases". In: *Science* 185.4157 (1974), pp. 1124–1131. DOI: 10.1126/science.185.4157.1124.

[6] Gerd Gigerenzer, Peter M. Todd, and ABC Research Group. *Simple heuristics that make us smart*. New York, USA: Oxford University Press, 1999.

[7] Herbert A. Simon. *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*. 4th ed. New York: Free Press, 1997.

[8] International Organization for Standardization. *ISO 31000:2018, Risk management—Guidelines*. Standard. Geneva, Switzerland: International Organization for Standardization, 2018. URL: https://www.iso.org/standard/65694.html.

[9] Department of the Air Force. *Risk Management Guidelines and Tools*. Tech. rep. DAFPAM 90-803. Department of the Air Force, Mar. 2022.

[10] Air Force Test Center. *Test Safety Program*. AFTCI 91-202. U.S. Air Force. Nov. 2022. URL: https://static.e-publishing.af.mil/production/1/aftc/publication/aftci91-202/aftci91-202.pdf.

[11] Department of the Air Force. *DAFI 90-802, Risk Management*. Tech. rep. Published on the Air Force e-Publishing website. Headquarters, Air Force Safety, Sept. 2024. URL: https://static.e-publishing.af.mil/production/1/af_se/publication/afi90-802/dafi90-802.pdf.

[12] Michael Stamatelatos and Homayoon Dezfuli. *Probabilistic Risk Assessment Procedures Guide for NASA Managers and Practitioners*. Tech. rep. NASA/SP-2011-3421. Washington, DC: NASA Headquarters, 2011.

[13] E. T. Bell. *The Development of Mathematics*. 2nd. New York: McGraw-Hill Book Company, 1945.

[14] Alan Hájek. "Interpretations of Probability". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2023. Accessed on 29 Aug 2025. Metaphysics Research Lab, Stanford University, 2023.

[15] Bruno de Finetti. *Theory of Probability: A Critical Introductory Treatment*. 2 vols., English translation of *Teoria delle Probabilità* (1937). London: John Wiley & Sons, 1974.

[16] John Maynard Keynes. *A Treatise on Probability*. London: Macmillan & co., 1921.

[17] Frank P Ramsey. "Truth and probability". In: *Readings in formal epistemology: Sourcebook*. Springer, 1926, pp. 21–45.

[18] A. Clayton. *Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science*. Columbia University Press, 2021. ISBN: 9780231553353.

[19] H. Jeffreys. *The Theory of Probability*. Oxford Classic Texts in the Physical Sciences. OUP Oxford, 1998. ISBN: 9780191589676. URL: https://books.google.com/books?id=vh9Act9rtzQC.

[20] E.T. Jaynes and G.L. Bretthorst. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. ISBN: 9780521592710.

[21] N. Silver. *The Signal and the Noise: Why So Many Predictions Fail-but Some Don't*. Penguin Publishing Group, 2012. ISBN: 9781101595954.

[22] A. Gelman et al. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2013. ISBN: 9781439898208.

[23] C. Perrow. *Normal Accidents: Living with High Risk Technologies*. Princeton paperbacks. Princeton University Press, 1999. ISBN: 9780691004129.

[24] David L. Brumley. *"MAKING SAFETY HAPPEN" Through Probabilistic Risk Assessment at NASA*. Tech. rep. 20200001592. National Aeronautics and Space Administration (NASA), Mar. 2020. URL: https://ntrs.nasa.gov/citations/20200001592.

[25] F. Galton. "Vox populi". In: *Nature* 75.1949 (1907), p. 7.

[26] P.E. Tetlock and D. Gardner. *Superforecasting: The Art and Science of Prediction*. Crown, 2015. ISBN: 9780804136709.

[27] D. Gardner. *Future Babble: Why Pundits Are Hedgehogs and Foxes Know Best*. Penguin Publishing Group, 2012. ISBN: 9780452297579. URL: https://books.google.com/books?id=6qNPEAAAQBAJ.

[28] I. Berlin, H. Hardy, and M. Ignatieff. *The Hedgehog and the Fox: An Essay on Tolstoy's View of History - Second Edition*. Princeton University Press, 2013. ISBN: 9780691156002. URL: https://books.google.com/books?id=V2iYDwAAQBAJ.

[29] J. Surowiecki. *The Wisdom of Crowds*. Knopf Doubleday Publishing Group, 2005. ISBN: 9780307275059.

[30] Mohamad Hjeij and Arnis Vilks. "A brief history of heuristics: how did research on heuristics evolve?" In: *Humanities and Social Sciences Communications* 10.1 (Mar. 2023). DOI: 10.1057/s41599-023-01542-z. URL: https://doi.org/10.1057/s41599-023-01542-z.

[31] D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011. ISBN: 9781429969352.

[32] G. Gigerenzer. *Calculated Risks: How to Know When Numbers Deceive You*. Calculated Risks: How to Know when Numbers Deceive You. Simon & Schuster, 2002. ISBN: 9780743254236.

[33] F.H. Knight. *Risk, Uncertainty and Profit*. Hart, Schaffner and Marx prize essays. Houghton Mifflin, 1921. ISBN: 9781548563509.

[34] Andy Stirling. "Keep it complex". In: *Nature* 468.7327 (2010), pp. 1029–1031.

[35] Frank Plumpton Ramsey. "Probability and partial belief". In: (1961).

[36] Herbert A Simon. "A Behavioral Model of Rational Choice". In: *Quarterly Journal of Economics* 69.1 (1955), pp. 99–118.

[37] Gerd Gigerenzer and Daniel G. Goldstein. "Reasoning the fast and frugal way: Models of bounded rationality". In: *Psychological Review* 103 (1996), pp. 650–669. DOI: 10.1037/0033-295X.103.4.650.

[38] G. Gigerenzer. *Risk Savvy: How to Make Good Decisions*. Penguin Publishing Group, 2014. ISBN: 9780698151437.

[39] G.A. Klein. *Sources of Power: How People Make Decisions*. The MIT Press. MIT Press, 1999. ISBN: 9780262260862.

[40] Jacob Q. Robinson. "Intuitive Judgment and Strategic Decisions". DTIC Accession Number: AD1160094. Master of Military Art and Science Thesis. Fort Leavenworth, Kansas: Advanced Strategic Leadership Studies Program (ASLSP), Sept. 2020.

[41] Hugo Bottemanne. "Bayesian brain theory: Computational neuroscience of belief". In: *Neuroscience* 566 (2025), pp. 198–204. ISSN: 0306-4522. DOI: https://doi.org/10.1016/j.neuroscience.2024.12.003. URL: https://www.sciencedirect.com/science/article/pii/S0306452224007048.

[42] Robert Bain. "Are our brains Bayesian?" In: *Significance* 13.4 (2016), pp. 14–19.

[43] Robert E. Kass and Adrian E. Raftery. "Bayes Factors". In: *Journal of the American Statistical Association* 90.430 (1995), pp. 773–795. ISSN: 01621459, 1537274X. URL: http://www.jstor.org/stable/2291091 (visited on 09/09/2025).

[44] Amos Tversky and Daniel Kahneman. *Evidential Impact of Base Rates*. Technical Report TR-4. Supported by the Office of Naval Research under Contract N00014-79-C-0077, Work Unit NR 197-058.

Stanford, CA: Stanford University, Department of Psychology, May 1981. DOI: 10.21236/ADA099501. URL: https://apps.dtic.mil/sti/html/tr/ADA099501/index.html.

[45] Wikipedia contributors. *Monty Hall problem — Wikipedia, The Free Encyclopedia*. [Online; accessed 18-August-2025]. 2025. URL: https://en.wikipedia.org/w/index.php?title=Monty_Hall_problem&oldid=1306215589.

[46] Steve Selvin. "A problem in probability (letter to the editor)". In: *The American Statistician* 29.1 (1975), pp. 67–71. DOI: 10.1080/00031305.1975.10479121.

[47] Marilyn vos Savant. "Ask Marilyn". In: *Parade Magazine* (Sept. 1990). Question posed in a letter by Craig F. Whitaker., p. 16.

[48] Marilyn vos Savant. "Ask Marilyn". In: *Parade Magazine* (Dec. 1990), p. 25.

[49] Edwin T. Jaynes. "Prior Probabilities". In: *IEEE Transactions on Systems Science and Cybernetics* SSC-4.3 (1968), pp. 227–241. DOI: 10.1109/TSSC.1968.300117.

[50] I. J. Good. "When batterer turns murderer". In: *Nature* 375 (1995), p. 541. DOI: 10.1038/375541a0.

[51] Royal Statistical Society. *Royal Statistical Society concerned by issues raised in Sally Clark case (News Release)*. Royal Statistical Society. Statement issued October 23, 2001, regarding the misuse of statistics in the courts. Oct. 2001. URL: https://rss.org.uk/RSS/media/File-library/Membership/Sections/2020/Sally-Clark-RSS-statement-2001.pdf.

[52] Richard P. Feynman. "Cargo Cult Science". In: *Engineering and Science* 37.7 (1974), pp. 10–13.

[53] Grant Sanderson. *Bayes theorem, the geometry of changing beliefs*. Dec. 2019. URL: https://youtu.be/HZGCoVF3YvM (visited on 07/27/2025).

[54] Hoesung Lee et al. "Climate change 2023 synthesis report summary for policymakers". In: *CLIMATE CHANGE 2023 Synthesis Report: Summary for Policymakers* (2024).

[55] Rebecca Holliday et al. "Record-breaking May heat in the UK: contrasting the extreme temperatures of 2024 and 1944 using climate attribution". In: *Weather* (2024).

Appendix A: Derivation of Bayes Factor

The Bayes Factor is the likelihood ratio that assesses the relative weight of evidence, data, or judgment between two competing hypotheses. Sir Harold Jeffreys first introduced the concept in his 1939 book, *Theory of Probability* [19]. A more widespread adoption by statisticians was encouraged and emphasized by Kass and Raftery in 1995 [43].

We derive the Bayes Factor from Bayes' Theorem and describe its interpretation. We will see that the Bayes Factor emerges naturally from Bayes' theorem when we compare competing hypotheses. Noting the symmetry of probability between two states, A and B, we have

$$P(A \cap B) = P(A|B)P(B) = P(B \cap A) = P(B|A)P(A) \tag{A-1}$$

where $P(A|B)$ is the probability of A given B. Bayes' theorem naturally follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{A-2}$$

The relationship is depicted geometrically in Figure A-1.[31]

Expressing (A-2) in terms of the probability of observing a hypothesis, H, given some data or evidence D, we obtain Bayes' theorem in its standard form:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} = \frac{P(D|H)P(H)}{P(H)P(D|H) + P(H^c)P(D|H^c)} \tag{A-3}$$

where:

- $P(H|D)$ is the *posterior probability* of hypothesis H given data D
- $P(D|H)$ is the *likelihood* of observing data D given hypothesis H
- $P(H)$ is the *prior probability* of hypothesis H
- $P(D)$ is the *marginal likelihood* or evidence, which normalizes the probability to ensure it lies between 0 and 1 (this will prove problematic)

The marginal likelihood of observing the data, $P(D)$, presents a computational challenge because it requires integrating or summing over all possible hypotheses. However, we can elegantly sidestep this issue by considering the alternative hypothesis and examining the ratio of posterior probabilities.

Let H^c represent the complement of hypothesis H. We can also write Bayes' theorem for this complementary hypothesis:

$$P(H^c|D) = \frac{P(D|H^c)P(H^c)}{P(D)} \tag{A-4}$$

[31] inspired by Grant Sanderson's 3Blue1Brown video [53]

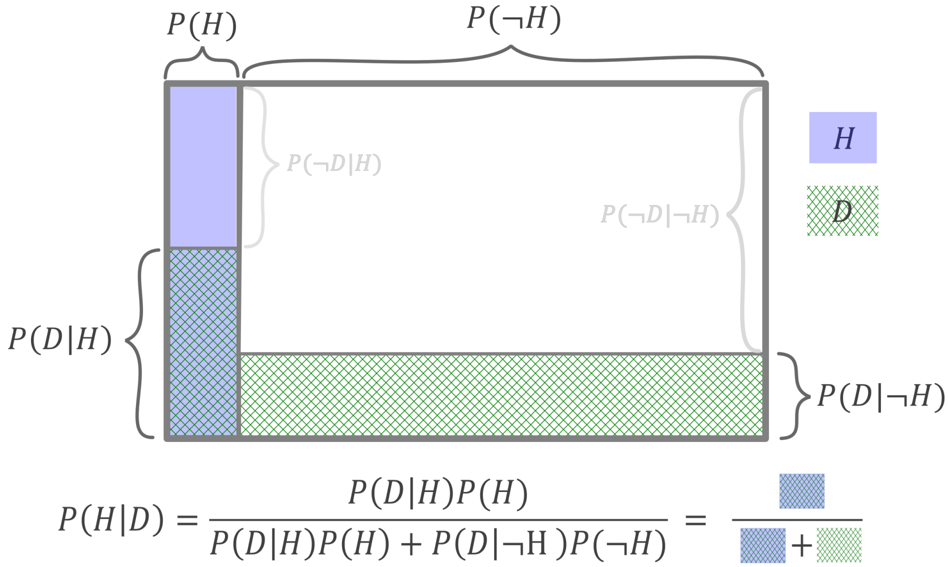


Figure A-1: *Geometric Interpretation of Bayes' Theorem*

The ratio of the two posterior probabilities, (A-3) and (A-4), is

$$\frac{P(H|D)}{P(H^c|D)} = \frac{\frac{P(D|H)P(H)}{P(D)}}{\frac{P(D|H^c)P(H^c)}{P(D)}} \tag{A-5}$$

Note that the problematic term assessing the probability of observing the data, $P(D)$, in (A-3)

$$P(D) = P(H)P(D|H) + P(H^c)P(D|H^c)$$

appears in both the numerator and denominator in (A-5), allowing us to cancel it. Thus,

$$\frac{P(H|D)}{P(H^c|D)} = \frac{P(D|H)P(H)}{P(D|H^c)P(H^c)} \tag{A-6}$$

This is the **posterior odds** for hypothesis H. The elegance of this formulation is that we have eliminated the troublesome marginal likelihood $P(D)$ entirely. Rewriting (A-6) to separate the likelihood ratio from the prior odds yields

$$\frac{P(H|D)}{P(H^c|D)} = \frac{P(D|H)}{P(D|H^c)} \cdot \frac{P(H)}{P(H^c)} \tag{A-7}$$

Defining the **prior odds** as

$$\Theta^{(0)} = \frac{P(H)}{P(H^c)} \tag{A-8}$$

and the **posterior odds**

$$\Theta^{(1)} = \frac{P(H|D)}{P(H^c|D)} \tag{A-9}$$

yields the definition of the **Bayes Factor**:

$$BF = \frac{P(D|H)}{P(D|H^c)} \tag{A-10}$$

Equation A-6 can now be rewritten in final form using the Bayes Factor:

$$\Theta^{(1)} = BF \cdot \Theta^{(0)} \tag{A-11}$$

Interpretation

The Bayes Factor, Equation A-10, represents the strength of evidence provided by the data D in favor of hypothesis H over hypothesis H^c. It quantifies how many times more likely the observed data is under hypothesis H compared to hypothesis H^c.

Interpreting the strength of the Bayes Factor:

- $BF \gg 1$: Evidence provides strong support for hypothesis H
- $BF > 1$: Evidence favors hypothesis H
- $BF \approx 1$: Evidence is equally consistent with both hypotheses
- $BF < 1$: Evidence favors hypothesis H^c
- $BF \ll 1$: Evidence provides strong support for hypothesis H^c

Combined Bayes Factor for N pieces of data is the product of the individual BF_i's so that

$$BF = \prod_{i=1}^{N} BF_i \tag{A-12}$$

The recent use of attribution methods in climate science [54, 55], which involves comparing the likelihood of particular climatic events using Monte Carlo simulations of two different models, with and without climate change effects, is an interesting example of the use of Bayes Factors.

Appendix B: Conjugate Prior Distributions in Bayesian Reasoning

A conjugate prior[32] is a prior distribution that, when combined with a particular likelihood function through Bayes' theorem, produces a posterior distribution from the same family as the prior with new parameters. This mathematical convenience allows us to compute our posterior analytically rather than resorting to complex numerical methods. This appendix summarizes two conjugate priors: the Beta and Gamma distributions.

Both the Beta and Gamma distributions are provided in readily available statistical software packages, making them straightforward to use.

- **Excel**: use the BETA.DIST(x, α, β, cumulative) and GAMMA.DIST(x, α, β, cumulative) functions; the boolean *cumulative* parameter returns either the CDF (TRUE) or PDF (FALSE)
- **SciPy**: Python's scipy.stats library contains beta.pdf (x, α, β), beta.cdf (x, α, β), gamma.pdf (x, α, β), and gamma.cdf (x, α, β)
- **NumPy**: Python's NumPy library contains np.random.beta (α, β, size) and np.random.gamma (α, β, size) for random sampling
- **MATLAB**: betapdf (x, α, β), betacdf (x, α, β), gampdf (x, α, β), and gamcdf (x, α, β) with corresponding random generators betarnd (α, β) and gamrnd (α, β)

Note that software packages may use different parameterizations, particularly for the Gamma distribution's scale vs. rate parameter.

Both the Beta and Gamma distributions are used as conjugate priors in Examples in Section IV.

B-I. Beta Distribution

The Beta distribution is a continuous probability distribution defined on the interval $[0, 1]$, which is a convenient conjugate prior for modeling proportions or Bernoulli/binomial trials in Bayesian statistics. The Beta distribution is parameterized by two positive parameters ($\alpha, \beta > 0$) which control the shape of the distribution. The probability density function is

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathcal{B}(\alpha, \beta)} \tag{B-1}$$

where $\mathcal{B}(\alpha, \beta)$ is the Beta function which serves as a normalization constant. The Beta function is defined as

$$\mathcal{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{B-2}$$

[32] conjugate comes from Latin *conjugatus*, meaning "joined" or "yoked together"

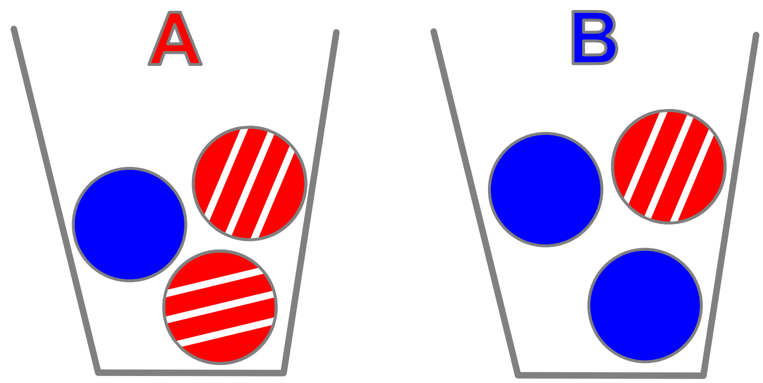


Figure B-1: *Two Bags with Three Balls Example*

where $\Gamma(\cdot)$ is the Gamma function:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt \tag{B-3}$$

The Gamma function satisfies the recurrence relation $\Gamma(z+1) = z\Gamma(z)$, so that it is often known as the 'factorial function,' owing to the property that

$$\Gamma(n+1) = n! \tag{B-4}$$

Table B-1: Statistics for the Beta Distribution ($\mathcal{B}(\alpha, \beta)$)

statistic	value	support
mean	$\mu = \frac{\alpha}{\alpha+\beta}$	$\alpha, \beta > 0$
mode	$\frac{\alpha-1}{\alpha+\beta-2}$	$\alpha, \beta > 1$
variance	$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	

The interpretation of the shape parameters is intuitive: α is the number of 'successes' and β represents the number of 'failures' in a series of Bernoulli trials (a binomial experiment). With a $\mathcal{B}(\alpha_{prior}, \beta_{prior})$ distribution, and x successes in n additional trials, the posterior distribution is also Beta-distributed and given by

$$\mathcal{B}(\alpha_{posterior}, \beta_{posterior}) = \mathcal{B}(\alpha_{prior} + x, \beta_{prior} + n - x) \tag{B-5}$$

Example: two bags with three colored balls

To illustrate the application of Beta distributions in Bayesian reasoning, consider the contrived academic example in which we have two bags, each containing three balls with two different colors:

- Bag A: 2 red balls, 1 blue ball (probability of drawing blue $= \frac{1}{3}$)

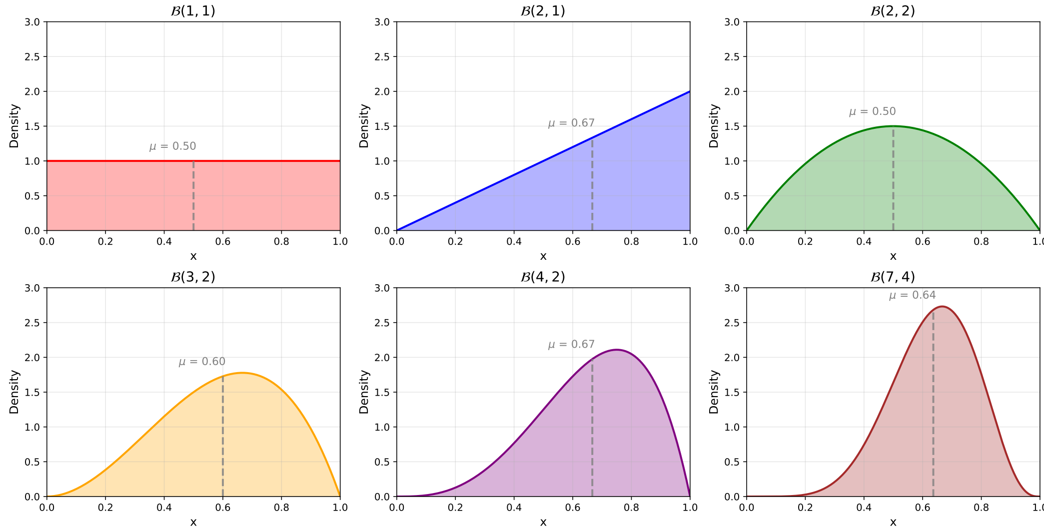


Figure B-2: *Posterior Beta Distributions for the Probability of drawing a Blue ball*

- Bag B: 1 red ball, 2 blue ball2 (probability of drawing blue $= \frac{2}{3}$)

One of the bags is selected at random, and we then draw balls from that bag (with replacement) recording the colors. We use Bayesian reasoning to update our prior belief about which bag was selected and what the probability of drawing a blue ball is using the Beta distribution as our conjugate prior.

Since a bag is selected at random, we have even our prior odds for selecting Bag B with two blue balls, thus $\Theta_B^{(0)} = 1:1$. Let's take the sequence of draws (with replacement) given in Table B-2.

Table B-2: Two-Bag Example

	draws	BF_i	$\Theta_B^{(i)}$	$p(B\|D_i)$	$\mathcal{B}_i$
0	—	—	$\Theta_B^{(0)} = 1:1$	$p = 1/2$	$\mathcal{B}_0(1,1)$
1	B	$BF_1 = \frac{2/3}{1/3} = 2$	$\Theta_B^{(1)} = 2:1$	$p = 2/3$	$\mathcal{B}_1(2,1)$
2	B,R	$BF_2 = \frac{1/3}{2/3} = 1/2$	$\Theta_B^{(2)} = 1:1$	$p = 1/2$	$\mathcal{B}_2(2,2)$
3	B,R,B	$BF_3 = \frac{2/3}{1/3} = 2$	$\Theta_B^{(3)} = 2:1$	$p = 2/3$	$\mathcal{B}_3(3,2)$
4	B,R,B,B	$BF_4 = \frac{2/3}{1/3} = 2$	$\Theta_B^{(4)} = 4:1$	$p = 4/5$	$\mathcal{B}_4(4,2)$
	$\vdots$				
9	B,R,B,B,R,R,B,B,B	—	$\Theta_B^{(9)} = 8:1$	$p = 8/9$	$\mathcal{B}_9(7,4)$

After a sequence of 4 draws consisting of 3 blue balls and 1 red ball, we are 80% confident that we had originally selected Bag B. The posterior distribution for the probability of drawing a blue ball is given in Figure B-2 for each six cases in Table B-2.

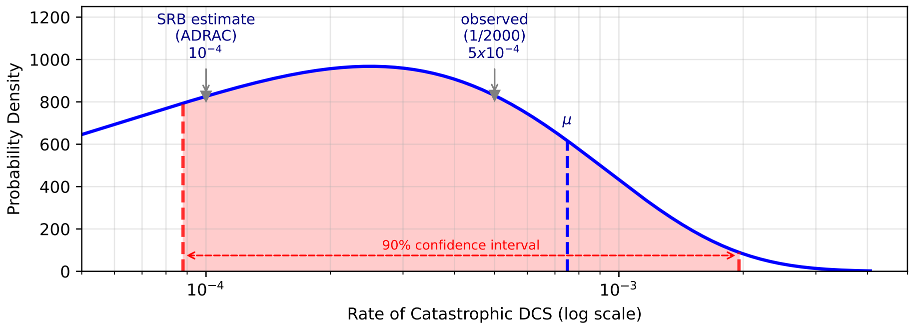


Figure B-3: $\Gamma(1.5, 2000)$ *Posterior Distribution for Expected Rate of a Rare Event*

B-II. Gamma Distribution

The Gamma distribution is a continuous probability distribution defined on the interval $(0, \infty)$ that serves as a conjugate prior for modeling positive-valued parameters such as rates, scales, and precision parameters. The exponential, Erlang, and chi-squared distributions are special cases of the Gamma distribution. The Gamma distribution is parameterized by two positive parameters: shape parameter $\alpha > 0$ and a rate parameter $\lambda > 0$. The probability density function is

$$f(x; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \tag{B-6}$$

where $\Gamma(\alpha)$ is the Gamma function defined by (B-3). An alternative parameterization uses the scale parameter $\theta = 1/\lambda$, yielding

$$f(x; \alpha, \theta) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-x/\theta} \tag{B-7}$$

Table B-3: Statistics for the Gamma Distribution, $\Gamma(\alpha, \lambda)$

statistic	value	support
mean	$\mu = \frac{\alpha}{\lambda}$	$\alpha, \lambda > 0$
mode	$\frac{\alpha-1}{\lambda}$	$\alpha > 1$
variance	$\sigma^2 = \frac{\alpha}{\lambda^2}$	

The Gamma distribution exhibits important conjugacy properties with several likelihood functions. For Poisson data with rate parameter λ, as was used in Section IV.H, a $\Gamma(\alpha_{prior}, \lambda_{prior})$ prior combined with observed count data $\sum_{i=1}^{n} x_i$ over n observations yields the posterior

$$\Gamma(\alpha_{posterior}, \lambda_{posterior}) = \Gamma(\alpha_{prior} + \sum_{i=1}^{n} x_i, \lambda_{prior} + n) \tag{B-8}$$

For exponential data with rate parameter λ, the Gamma distribution serves as a conjugate prior, making it useful for modeling failure rates and other positive, continuous random variables.