Preventing Premature Conclusions – Analysis of Human-In-the-Loop Air Combat Simulations

Matthew R. MacLeod
Centre for Operational Research and Analysis
Defence R&D Canada
101 Colonel By Drive, Ottawa, ON K1A 0K2
Matthew.MacLeod3@forces.gc.ca

August 2012

Abstract

Comparisons of fighter aircraft, and of the tactics used to employ them, are increasingly conducted with Human-In-the-Loop simulation. This has many advantages, but the implementation of probabilistic outcomes within air combat and other simulations leaves unwary decision makers vulnerable to multiple cognitive biases, which in turn can result in premature or erroneous conclusions – often reinforcing existing preferences. By consequence, the responsibility of the operational researcher participating in these trials is often not only to draw conclusions from data, but also to actively discourage the drawing of conclusions not borne out by the evidence. This paper covers tools and techniques an analyst can use to identify and convey to trial participants the impact of random variables on results, based on the author's experience with common pitfalls. There is a particular focus on practical methods that can be implemented on-the-fly, as trials and data are often subject to strict security controls that complicate post-trial analysis and subsequent communication of results. Key recommendations for decision makers include focusing on the robustness of a platform or tactics to inherently uncertain events (e.g. missile engagements), and defining risk levels in terms of acceptable probabilities of success. Potential mitigations via experimental design are also briefly discussed.

1 Introduction

Human-In-the-Loop (HIL) virtual trials are an attractive method for examining air combat, and the aim of this paper is to describe some of the challenges inherent in analysing the results to reach operationally significant conclusions. The advantages include: reduced expense in logistics and fuel, increased flexibility in threat simulation, ability to control (albeit simulated) environmental conditions,

and, not least, the ability to model future aircraft and aircraft capabilities. Given the ability to model current and future adversary threats, as well as future domestic capabilities, two natural applications emerge that are distinct from what is possible in live exercises:

- developing and evaluating tactics to be used against current and future threats that cannot be well simulated (in the physical sense) by friendly aircraft; and
- evaluating or comparing future friendly aircraft or aircraft capabilities.

It is important to note the emphasis on evaluation and comparison in the above applications, and consider also the impact of the decisions being made. It is incumbent upon the operational analyst to ensure that the conclusions drawn at these trials are valid and unbiased – which more often than not involves demonstrating why certain conclusions cannot be drawn, whether due to insufficient evidence or otherwise. This paper will draw on the author's experience and the statistical and psychological literature to present practical advice for the analyst providing decision support in HIL air combat trials.

To facilitate discussion, some terms will be briefly defined. A *run* will refer to a single instantiation of the air combat simulator, with the aircraft fighting until they reach the limit of their fuel, weapons, or one side is eliminated. A *trial* will refer to a sequence of runs, generally conducted over the course of a week or weeks. A *trial director* refers to the individual, often the ranking military member, with authority to set the conditions of each run, define mission success, and declare runs over. An *analyst* refers to an operational analyst or researcher. Canada's operational researchers are developed to be generalists, and so the approaches described herein are written with the generalist in mind, rather than the air combat specialist.

In a virtual simulation, physical processes that are not modelled directly must be ignored, treated as deterministic, or have their outcomes determined stochastically. The author's experience is that trial directors have an understandable desire to reduce the number of independent *variables* they are examining given limited time and other resources, and prefer to assign all other variables some fixed value. This can be deceptive, however, in the case of random binary variables such as missile end-game outcomes. As their probability function can be fully defined by a single value (the probability of success p), it can seem that by setting p the variable has been given a 'fixed' value, and therefore can be removed from consideration.[1] Much of this paper will deal with the recognition and communication of the *variability* which follows from giving what appears to be a fixed value to what is in fact a *variable*.

[1]The author will note that as simulators advance p may not be a single value, and are starting to incorporate aspect angle and other factors. While adding realism, this makes direct comparisons of complex engagements much more difficult.

That being said, although reducing variability may be desirable from a purely analytic perspective, for HIL simulations one must consider the human factors implications of following a structure that is overly predetermined. If the behaviour of an opponent or the outcome of a missile launch is too predictable, for instance, participants may become bored. Or, even worse, they may adjust their own behaviour to "fight the simulation," rather than behaving in a naturalistic manner. This is particularly the case if they do not feel the need to conduct realistic battle damage assessment, and therefore turn away from an engagement sooner than they otherwise would. The difficulty in balancing the implied requirement to present an unpredictable and challenging scenario to the pilots against the need to evaluate a predetermined set of variables is at the root of the problem explored in this paper.

There is another important reason for considering variability, in that it is inherent to modern fighter combat, and is not an artifact of the simulation. As individual aircraft sensor capability has increased, the need for large numbers of aircraft in formations has arguably decreased. This has also been reflected in declining fighter fleet sizes. The consequence is that most air-to-air encounters will be well into the realm of small sample sizes, where standard intuitions and rules of thumb based on the law of large numbers do not apply, and one must be careful of relying on expected values. For comparison, one can consider the analogy of a few on few air combat encounter to a few on few sporting event. Even if one knows the strategies, the players, and the equipment that either side will use, one can generally only hope to know the *expected* outcome. The author would assert it is not the intent of most nations to conduct a war of attrition with their tens of expensive fighter aircraft, and so the average outcome – often expressed as an "exchange ratio" – is less relevant than characterizing the *risk*[2] of failure in a given scenario. Increasing the number of runs to narrow the confidence interval of results cannot 'fix' the issue; the more important goal is to explore the space of possible outcomes, to determine and mitigate any potential sources of failure. One approach is to consider both the number of unthreatened shot opportunities, as well as the total number of missiles available to the friendly formation, and to apply the expected distribution of success for those missile firings in post analysis. The issues with expectations in this context will be explored in Section 2. Cognitive biases with respect to the likelihood of streaks in small samples will be discussed in Section 3. Additional biases and statistical issues with making comparisons between runs will be examined in Section 4

Against that background, implications to practice and practical methods for mitigating the issues raised will be presented in Section 5. A further complication for most air combat analysis is that nations closely guard studies of the relative effectiveness of fighter aircraft. This imposes tight restrictions on where the data

[2]Or, more precisely, the likelihood of failure, with risk being defined as the likelihood times the consequence.

can be generated and analyzed, and also on its later distribution. The practical effect of these restrictions are large delays in data dissemination. As such it is the author's experience that simple analysis delivered *in situ* is much more useful than a detailed follow-up analysis. The creative challenge for the analyst is to assertively and clearly convince a notoriously cocksure group of individuals (i.e. fighter pilots) that their intuitions are wrong, in a limited amount of time, with limited tools, and with sometimes limited access to the underlying data. The aim of this paper is not to provide guidance on the most complete analysis, but to suggest to analysts strategies for maintaining the utility and scientific integrity of results developed on-the-fly in a challenging environment.

Finally, the advice presented will be summarized in Section 6. This will take three forms: mitigating strategies in trial design, advice to the analyst, and advice to trial directors.

2 Randomness and Expectation

Virtual simulations are simplifications of the natural world, at least for practical purposes. Physical processes that are not modelled directly must be ignored, treated as deterministic, or have their outcomes determined stochastically. Deterministic outcomes may make sense for certain processes that have a reasonably well defined best case or worst case – e.g. detection ranges – depending on one's objectives. This is not always practical. In particular, outcomes that are in reality binary with non-extreme probabilities of either event cannot generally be assigned to always have one or the other outcome. One could try to treat them deterministically by re-running the trial with both values, but it is easy to see that as the number of such outcomes n rises, the total number of trial runs required quickly becomes impractical: rising as 2^n. The most practical option therefore appears to be to limit the deterministically set outcomes to those assumed to have the greatest effect of interest, and for the rest either ignore them or treat them stochastically.

As stated in the introduction, trial directors will generally attempt to reduce number of independent *variables* they are examining, and so assign other variables some fixed probability. For binary variables, this function is really a single value, the probability p of the variable taking one of its values (the other necessarily taking probability $q = 1 - p$). The author's contention is when picking a fixed p, the director assumes they are giving the associated variable a fixed value, thereby removing it from consideration. While this may be true when the law of large numbers applies – and the effect will in fact be averaged out over time – it often does not apply to processes of great import to air combat simulations.

Analysts and pilots alike have been heard to lament that 'statistically valid' conclusions aren't possible at a given trial because of the small number of runs. As raised in the introduction, however, this in some sense misses the point –

small sample sizes are the reality of modern air combat. What is perhaps more attainable is to look for a tactic that 'almost always' works, or for which at least the risk of failure can be quantified. This section will focus on what, in the author's experience, is the dominant stochastic factor.

The loss of even a single aircraft in a few on few encounter (particularly out of two or four total) is clearly significant by any measure. Although losses may occur due to e.g. an aircraft running out of fuel, by and large it is the effect of missile firings that determines losses. Conversely, the number of air-to-air weapons being carried by a given aircraft may be limited by a requirement to carry weapons internally, to carry a mixed load of air-to-air and air-to-ground weapons, or to do both.[3] As a 'small' (this definition will be refined) number of weapons is being used against a 'small' number of aircraft, the variance may be 'large' with respect to the average under conditions of interest.

It is important to understand what those conditions of interest are. In general, one will only invest the time and money in running an HIL simulation if the fight is somewhat fair – i.e. that the outcome is in doubt. This should not be read to mean that the encounters being simulated are predicted to be typical, it simply means they are the cases that require the most study. In reality, there will be many possible encounters which are not worth much, if any, simulation time, because one side has a clear advantage in engagement range. Given that one will only be simulating cases with some uncertainty as to the survival or destruction of aircraft on each side, the dominance of missile outcomes will become clear.

To discuss missile outcomes in a simulation, it is important to define what factors are captured directly and indirectly by the model. The overall Probability of Kill (P_K) (i.e., the likelihood of target destruction) given a trigger pull comprises factors including:

Probability of Launch (P_L) is the probability that the missile will fire successfully. This is relatively easy to simulate using a Pseudo-Random Number Generator (PRNG), but is often set to 1 or rolled into another probability.

Probability of Intercept (P_I) is the probability that a successfully fired missile will intercept a target. This factor depends on the relative position and velocities of the aircraft at launch, as well as the subsequent manoeuvres of the target. One of the primary advantages of virtual simulation is that the kinematics of missile fly-out can be directly simulated. For that reason the P_I is often estimated as a result of a trial, rather than being an input. It may or may not include the effect of any countermeasures employed by the target.

End-game P_K is the probability that the missile will destroy the target, given that it intercepts. In physical encounters this depends on the kinematics

[3] Standard configurations for the Eurofighter Typhoon, for instance, include either four or six Beyond Visual Range (BVR) missiles [1].

at end-game, the missile fusing and warhead, and last ditch manoeuvres by the target. These are computationally expensive to compute in a live simulation, and so are generally rolled into a single probability value evaluated against a PRNG.[4] Depending on the simulator, factors such as P_L and countermeasures may also be considered in determining the probability value. In the author's experience this is normally what simulator operators mean when they say P_K, so care must be used to understand the difference between the physical meaning of P_K and the end-game only process used in the simulator.

The important thing to note is that whether one is referring to overall P_K or to end-game P_K, there are two possible outcomes: the target is destroyed or not. The discussion below regarding binomial distributions therefore applies in both cases – the starting number of missiles will just be lower if one is considering only those that have reached end-game. The variance of a binomial variable over a small number of trials will be seen to be large, particularly when considering the small sample sizes – establishing the need to account for this in both simulation and in tactics development.

To give the reader a sense of the potential spread in results, Table 1 gives the expected number of kills, K, and a spread of two standard deviations, σ, for different probabilities, p,[5] and two possible total missile loads for a formation of aircraft.[6] The use of standard deviations as a measure will be further discussed below. The widest spread is for 24 shots at $p = 50\%$: one could reasonably expect to achieve anywhere from seven to seventeen kills. Even on one of the narrowest cases, $p = 90\%$, anywhere from twelve to sixteen kills are within the realm of expectation. A range of expectation of ten – or even just four – kills in a few on few engagement is a major factor. Conversely, if one is planning missile loads based on expected missiles to be used per target, it represents a wide spread when considering efficiency and the competing uses for weapons stations (e.g. other ordnance, fuel tanks); there is also risk on the 'lucky' end of the spectrum when operating a Short Take-Off Vertical Landing (STOVL) aircraft with limited vertical bring-back weight. It also means that comparing the relative effectiveness of aircraft or tactics between runs must be undertaken with caution; a difference of two or three kills between two runs may seem significant, but in reality is likely well within statistical expectations on purely random factors.

Accepting that the stochastic outcome of missile firings induces a significant

[4] As noted in an earlier footnote, this is beginning to change, and aspect dependence is becoming more common. The same general issues will apply, although the analysis of what constitutes an 'expected' result will become non-trivial.

[5] Missile P_K are highly sensitive information. To avoid any issues with assuming that a specific missile/aircraft is being referred to, a symmetric spread of possible probability values is given here and throughout. The reader is free to interpolate or calculate their own exact values.

[6] 24 missiles could represent a formation of four aircraft with six missiles each, or six aircraft with four each. 16 missiles could represent four aircraft each carrying four.

	$E[K] \pm 2\sigma$	
p	24 shots	16 shots
10%	2.4 ± 2.9	1.6 ± 2.4
25%	6 ± 4.2	4 ± 3.5
50%	12 ± 4.9	8 ± 4.0
75%	18 ± 4.2	12 ± 3.5
90%	21.6 ± 2.9	14.4 ± 2.4

Table 1: Expected number of successful missile kills K out of 24 and 16, for several values of p.

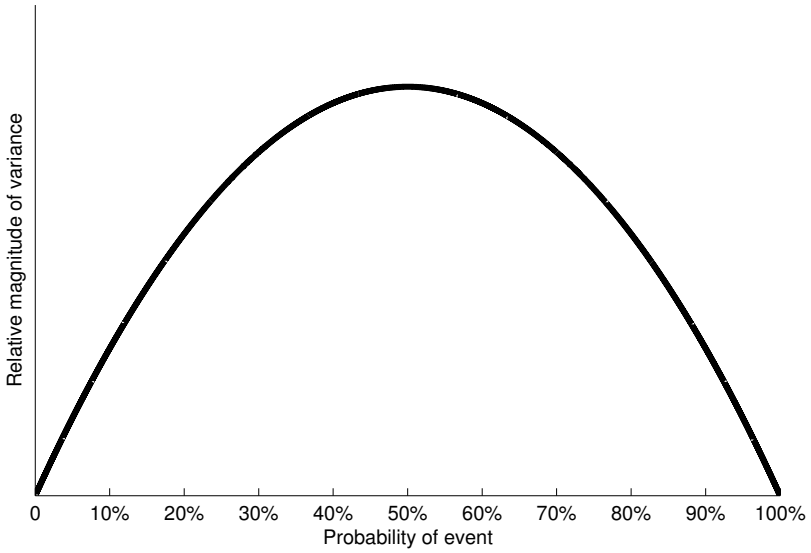

Figure 1: Magnitude of variance for a binomial distribution given the probability of either event, on a linear scale.

amount of dispersion on the results of air combat encounters, one can then move on to understanding how to communicate the effects to trial participants. A first point to consider is that, all else being equal, the absolute value of σ will be higher the closer p is to 50%. That is to say, the more 'fair' a coin flip is, the higher the uncertainty of the result. This relationship is visualized in Figure 1. This also means that the higher a P_K is above 50%, the smaller the variance will be – an additional benefit to improved performance. Standard deviations are particularly easy to calculate for binomial distributions ($\sigma = \sqrt{np(1-p)}$), and so are an easy and quick communication tool in dynamic, time-limited situations.

Before continuing, it is important to place a few caveats on the interpretation of standard deviations in this context. Although the binomial distribution is often well-approximated by the normal distribution, this does not hold for small sample sizes. Therefore common rules of thumb regarding ~68% of results falling within 1σ of the mean, μ, and 95% of results within 2σ, do not hold. In fact, due to the discrete nature of possible results, it is not generally possible to find a range of values around μ that will contain any given pre-specified percentage of the results. The *skew* (asymmetry) of the distribution also becomes a factor as p moves away from 50%. For p greater than 0.5, values outside of two standard deviations are more likely to be above the mean than below, and the opposite for p less than 0.5; that is to say the good news gets better for missiles with higher P_K, and the bad news gets worse for very low values. All that said, standard deviations are easily calculated without special tools, do not have to be quantized, and they do have at least a weak lower bound[7] on the fraction of the probability density function they contain. They are therefore considered to be a useful way to communicate the effects of dispersion, as they have a well-defined meaning, which at worst will not be too far from their vernacular meaning.[8]

To summarize, the outcome of missile firings is a major factor in air combat. Accepting that the outcome of each firing is a binomial variable, the variance at small sample sizes introduces a significant amount of uncertainty to the results of an encounter. This is certainly true for simulations using a PRNG to evaluate the outcome of each shot. Communicating this uncertainty, and how it limits the ability to compare the results of different runs, will be developed further in the remainder of this paper. In real encounters, the exact value of p will likely be unknown, and will be different depending on the exact conditions at firing. However, the results will still be subject to uncertainty in a manner approximated by the binomial distribution. The point to be conveyed is that

[7] Chebyshev's inequality specifies that no more than $\frac{1}{m^2}$ of a distribution's values can be more than $m\sigma$ away from μ. For example, at least 75% of the values of a distribution must be within two standard deviations.

[8] For fuller discussions of the issue of confidence intervals for small sample sizes with the binomial distribution, see e.g., [2],[3]. If one has more time to develop tools for a specific trial, they contain more advanced approaches to consider.

planning tactics and strategies using only the expected ('average') results of missile firings is, at best, incautious. Uncertainty must be understood as risk.

3 Streaks

> "The reliance on heuristics and the prevalence of biases are not restricted to laymen. Experienced researchers are also prone to the same biases—when they think intuitively."
>
> Amos Tversky and Daniel Kahneman [4]

Both the general population and trained professionals are prone to several cognitive biases with respect to statistical thinking. In particular, humans have been shown to consistently underestimate the likelihood of streaks in random processes. This has been attributed to an apparent belief that short sequences should be as representative of their generating processes as long ones [5, 6]. This expresses itself in various forms, including the "gambler's fallacy" and an unjustified belief in "hot hands" in basketball.

For a fair coin, for example, the probability p that the result of the coin flip will "alternate" (i.e. be the opposite of the last flip) is 0.5. However, when shown sequences with various alternation probabilities and asked to judge whether they were produced by a fair coin, subjects rate those with a 0.7–0.8 chance of alternation as being the most 'random' [6]. That is to say, people's intuitive judgement is that short sequences or random events will 'even out' substantially more quickly than they will in practice. This is commonly known as the gambler's fallacy.

A seemingly opposite misperception is the the hot hand fallacy (see [6]), most commonly seen in sports. In those situations, humans tend to believe that a player on a streak is likely to continue that streak. The key distinction drawn is that humans see a coin flip as random, whereas they see an athlete as having agency over events – i.e. a human can be "hot," whereas an inanimate object cannot [7]. It has been shown that this perception can be manipulated: when an observer's attention is focused on the person flipping a coin, rather than the coin itself, predictions reflect the gambler's fallacy less, and look more like the hot hand fallacy [8]. It can be argued that a missile outcome has aspects of both situations – people may mentally assign the outcome of a shot to either the missile or the person firing it. In any case, there is ample cause for concern that the frequency of streaks will be misperceived.

This becomes an issue in air combat tactics development, in that the author has frequently observed that pilots underestimate the likelihood of successive misses (they may be more willing to attribute misses to the inanimate missile than to themselves). They may as a result develop tactics that work assuming that the outcome of a small number of missile firings is representative, while discounting the cases where a run of 'bad luck' may leave them in a less favourable

position. It behooves the analyst to be prepared to calculate the probability of an 'unlikely' series of misses, and to challenge the air crew on the robustness of their tactic.

The probability of having n successive misses in a row is easily calculated on the fly: $(1-p)^n$. This can easily be done if a particular pilot complains about the 'black cloud' over their head. With a bit more work, one can calculate what is perhaps more telling: the probability of *not* having a streak of a certain length over the course of an encounter. Figure 4 shows examples for runs where 24 and 16 shots are taken, respectively. Some key points are worth drawing attention to: to have a greater than even chance of not seeing a streak of more than three misses in a row in 24 shots, one must have a P_K higher than 65%. Even at $P_K = 75\%$, there is a close to 24% chance of seeing a streak of three consecutive misses, and almost 7% chance of seeing a streak of four. For the same P_K for 16 shots, there is still a $\sim 16\%$ chance of seeing three consecutive misses, and almost 4% chance of seeing four. If one's tactic relies on *not* having that many misses in a row, these *chances* become *risks*.

The overall message of this section in that even if the P_K of a missile is 'high,' a tactic still needs to be robust against streaks of successive misses. This can be conveyed in the language of risk, as in the probability of not having a streak of a certain length. This is of particular concern if the streak happens to one or two isolated aircraft.

4 Comparisons

> "That's just not cost effective, especially since in the end this is not a scientific survey. It's a random survey."
>
> Daniel Webster, US Congressperson, on sponsoring a bill to eliminate the American Community Survey [9]

For this section, we first return to the discouraging discoveries of Tversky and Kahneman:

> Our thesis is that people have strong intuitions about random sampling; that these intuitions are wrong in fundamental respects; that these intuitions are shared by naive subjects and by trained scientists; and that they are applied with unfortunate consequences in the course of scientific enquiry.
>
> We submit that people view a sample randomly drawn from a population as highly representative, that is, similar to the population in all essential characteristics. Consequently, they expect any two samples drawn from a particular population to be more similar to one another and to the population than sampling theory predicts, at least for small samples.

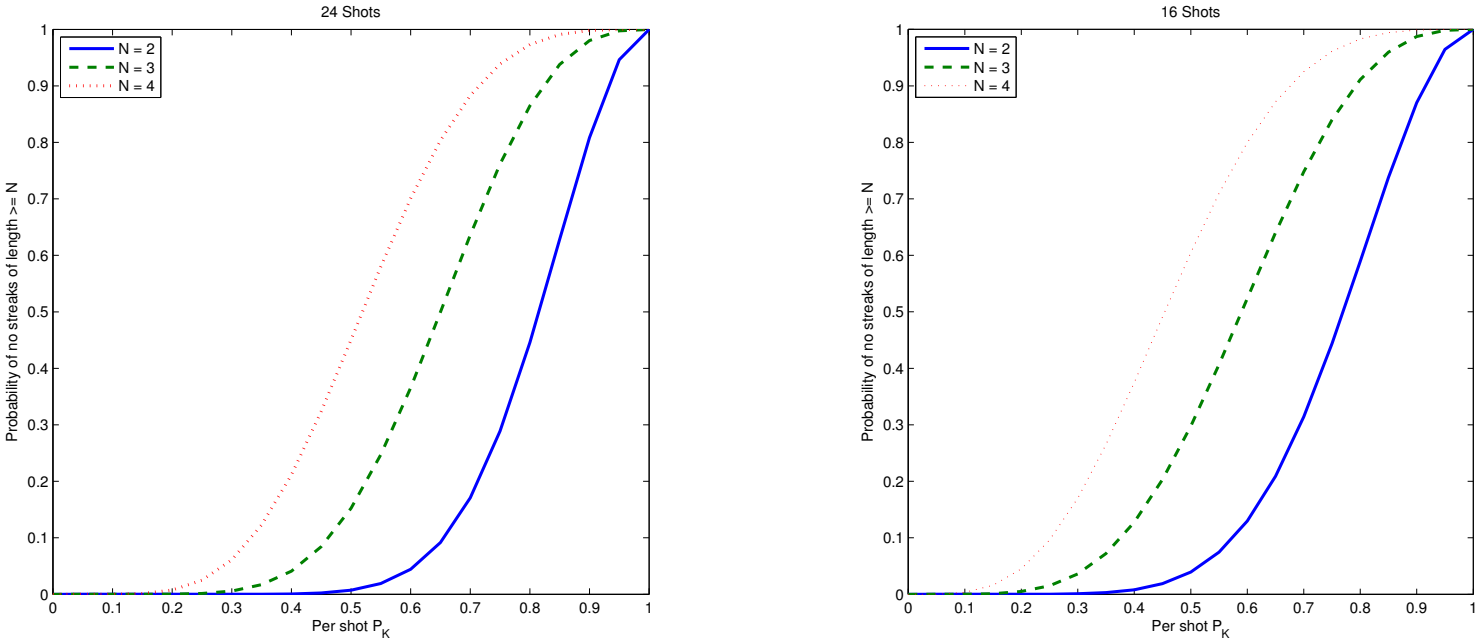

Figure 2: Probability of not having a miss streak of length 2, 3, or 4, in 24 or 16 successive shots, given a P_K

...

> In the absence of a statistical test, our respondents followed the representation hypothesis: as the difference between the two samples was larger than they expected they viewed it as worthy of explanation. However, the attempt to "find an explanation for the difference between the two groups" is in all probability an exercise in explaining noise. [5]

It is the author's experience that these misleading intuitions tend to dominate conversations at simulation events – even when both pilots and analysts are actively aware of the issue. Statements may start with comments like "well, it's not statistically valid, but..." As the reference above shows, this is not a fault that should be attributed to individuals, but a powerful effect due to the way the human brain works.

As a consequence, the responsibility of the analyst in these events is often not (or not only) to draw conclusions from the data, but to *actively discourage the drawing of invalid conclusions*. The analyst is usually not the primary designer of the trial, and in any event so many variables are present that it is essentially impossible to draw quantitative conclusions in a reasonable amount of runs. However, there are runs that can be identified as clear outliers, which may not be immediately obvious to the participants. The analyst may therefore provide value by:

- preventing the premature acceptance of the superiority of a combination of Tactics, Techniques and Proceduress (TTPs) because of one 'lucky' run;
- preventing the premature rejection of a combination of TTPs because of one 'unlucky' run.

It is the author's experience that the second role is in fact more important. TTPs that are seen as successful will tend to be incorporated and built upon in subsequent runs, and flaws are more likely to become apparent. One 'unlucky' run may prematurely condemn a novel idea to the scrapheap – particularly if it is unpopular. Where strong personalities come into play, a few 'unlucky' rolls may be used to explain away failure, whereas 'lucky' rolls are almost never pointed out as a reason for success. With all that being said, when different TTPs are being compared based on two runs on either side of the probability distribution, clear presentation of that fact can play a role in encouraging the group to collect more data before coming to a conclusion.

4.1 Comparison of Successive Runs and Regression to the Mean

Given the difficult-to-combat tendency to attempt to explain any difference between runs – even when one knows that the difference is well within statistically

expected variation – there is an additional effect of which it is important to be aware. *Regression to the mean* is a phenomenon that at first may seem contradictory to the notion of independence of events, but is in fact a consequence of statistics [10].

This concept can be explained (see e.g. [4]) by imagining a group of 100 students who are given two equivalent aptitude tests. If one looks at the top ten performers on the first test, one will find that they will have – on average – done slightly worse on the second test. Similarly, if one looks at the bottom ten performers from the first test, they will on average have done better on the second test. What is important to note, however, is that this will be true whichever test one considers to be the 'first' and 'second.' That is to say, it is not that all students will tend to regress toward the mean from one test to the other, it is that the way one selects the students has statistical implications. Not every top performer will do worse, and not every low performer will do better, it is the change in the average of each group that is predictable.

The underlying mechanism is that there is almost always some random component to how a person performs on a given test – e.g. specific questions that they have studied may or may not appear, they will get a few questions right or wrong by guessing, etc. The highest performers on a given test will tend to have been maximally 'lucky,' or somewhere to the right of their mean performance on a probability curve. Although for each individual top performer there is still some chance they will do even better on the next test, the greater mass of the probability curve is below their first value, and it is more likely that they will do worse. We can therefore expect that the *average* performance of groups of extreme performers will revert toward the mean. It is important to note that this applies just as much if the analysis is done 'backward in time,' although that case is less relevant to the discussion of cognitive biases.

As a more practical example of the implications of this phenomenon, the authors of [4] observed a failure to appreciate this fact in a group of flight instructors. The instructors had developed an intuition that when they praised their students for exceptionally good landings, the students tended to do worse on the next attempt. Similarly, when they criticized them after a poor landing, they tended to do better. The instructors had therefore concluded that criticism was more effective than praise, contrary to psychological studies on the subject. However, if they had refrained from offering any feedback, the instructors would have observed the same regression toward the mean.

This effect is important to appreciate in the context of air combat simulation, in that different tactics are often compared based on successive runs. If a run is particularly successful due to random factors (including successful missile end-game dice rolls), whatever is attempted in the next run is likely to turn out less well. Similarly, whatever is attempted after a particularly 'unlucky' run will likely work better. The tacticians must therefore resist the temptation to read too much into successful or unsuccessful runs, and even more so the runs immediately following. While in a cold quantitative sense this may not seem

relevant, in a situation where decisions are being made *in situ* as to how to proceed, the real qualitative effects on the perceptions of the participants should not be discounted.

Given this phenomenon, combined with Section 2 one needs to be very careful when judging the relative merits of tactics or aircraft. The author would advocate instead a pass/fail mindset for evaluating a given combination of factors at a given level of risk. For instance, instead of looking at end result including all the stochastic factors, one may focus on how many unthreatened shot opportunities can be obtained given the performance of one's missiles and aircraft. When then combined with the likelihood of a certain number of successive failures due to the P_K, one can obtain a quantitative measure of the risk of a particular tactic in a particular set of circumstances.

4.2 Ordering of Outcomes

The discussion thus far has disregarded the impact of the order of outcomes, in order to simplify the discussion and analysis. However, this is rarely, if ever, a valid assumption. Perhaps most clearly, the probability of success will change if unfired missiles are destroyed along with their host aircraft thanks to an early successful kill. Furthermore, the context of the missile engagement will also impact the overall result – a kill against a well-positioned and threatening aircraft will help the formation more than one against a retreating opponent.

Considering the permutations of outcomes as well as their likelihood of occurrence further increases the difficulty and depth of the trials and analysis required. However, if they are left out, one must be consciously aware that they are leaving out potentially critical differences between runs. One may view the discussion herein as a first step toward increasing the understanding of trial participants, with the potential for further expansion.

4.3 Other Random Effects

Although it has been argued that the large variability in outcomes created by missile effects will tend to swamp other effects, they are certainly not the only random effect that can have a disproportionate effect. Both in real encounters and in many models, detection by radar, infrared, and other systems are not completely deterministic. Small changes in relative aspect may also be the difference between whether an aircraft is detected or not – and following initial detection, tracking and subsequent engagement become much easier.

Individual tendencies of pilots can also effect the outcome in this way, as their tendency to fly a little faster or slower, to turn with greater or lesser forces, and to manoeuvre and employ counter-measures with maximum effectiveness can and does vary. This is a particular factor when humans are flying both sides of the engagement, as individuals may not have achieved great proficiency on in particular the aggressor systems. It is therefore worth repeating that the

discussion herein is a first step, and that there are further effects to be teased out as analysis deepens.

5 Implications to Practice

Cognitive biases are real and unavoidable, both for pilots and analysts, and therefore must be managed. As Tversky and Kahneman discovered, "[a]pparently, acquaintance with formal logic and with probability theory does not extinguish erroneous intuitions" [5]. The analyst must be ready and prepared to deal with sentences that begin "it's not statistically valid, but..." on a routine basis.

It is also important to remember that the variance of small sample sizes are an issue in real combat, not just in simulation – simply running longer trials to narrow the confidence intervals does not fix the problem of real risk to real tactics. The relevance of broad averages such as aircraft exchange ratios and missile per target pairings to modern air combat must be questioned, particularly when presented without any measure of confidence interval. It is the risk of failure that is the real issue to be quantified.

5.1 Examples of Displaying Probabilities

This subsection will present a number of possible presentations of probabilities and variances to observers, specifically the trial director and mission leads. The most effective depends on the personality and learning styles of the individual observer, and so none is recommended as the definitive solution.

The most simple is presented in Table 2. This is trivial to implement in whatever spreadsheet program is available in the environment in which the simulation is held,[9] even if it must be quickly built from scratch. This is particularly useful when P_I is considered to be separate from end-game P_K – one can quickly convey to the trial director that *even if* the formation is able to guide all of its remaining missiles to intercept, there is a certain Probability of Success (P_S) of killing the remaining targets. If this is sufficiently low, this may instigate the termination of the run, allowing a hopefully more productive run to be started.

P_K	90%
Shots	16
Targets	12
P_S	98.3%

Table 2: Simple dynamic calculator display for probability of success, given number of remaining missiles and targets.

[9] In Microsoft Excel, the function in the last cell is:
`BINOMDIST(Shots-Targets,Shots,1-P_K,TRUE)`.

Particularly if the number of targets and missiles is going to be fixed over a number of runs, the simple dynamic calculation can be expanded out into tabular form, as displayed for several illustrative values of P_K in Tables 3a, 3b, and 3c. The current P_S can then be traced through the table as the encounter develops. This can be unwieldy, however, depending on the number, size and location of displays in the simulation facility. For inclusion in tactical recommendations, where the number of missiles to be carried for a given encounter is given, and the number to be employed in a given situation is also being presented, the associated probabilities of success can also be pre-calculated and listed alongside.

For the more visually-minded, displaying the spread of standard deviations over the course of an engagement can be helpful. Examples are given in Figure 3,[10] for $P_K = 0.75$, $P_K = 0.5$ and $P_K = 0.25$. A 2σ spread is shown as a coloured region about the mean, μ with respect to the *vertical* axis, and clipped to physically possible outcomes (i.e. if $\mu + 2\sigma$ is greater than the initial number of targets, the region is not extended above that). In each case the initial number of targets is set to the average number of targets that could be destroyed with 16 shots, so it should be noted that the height of the range at 0 shots remaining for $P_K = 0.75$ and $P_K = 0.25$ is actually the same, and less than that for $P_K = 0.5$, in accordance with Figure 1. Example random walks through the space are presented as green (successful), yellow (successful, but with no left over missiles), and red (unsuccessful) lines.

What is immediately apparent in this visualization is the two directions in which outcomes can vary from the expectation, and their significance in physical terms. With $P_K = 0.75$ it is not unexpected that – all else being equal – the encounter could finish with three targets remaining, or with four missiles left over. This is even more dramatic when $P_K = 0.5$; it would be just as unsurprising that four targets could escape, or that the aircraft could have only needed ten of sixteen missiles to destroy the eight targets. These are clearly quite different outcomes, and can quickly give tacticians the sense of the range of possibilities for which they need to be prepared.

This display can also be quite useful when pilots have come off an apparently quite successful mission that in reality was due to random chance taking them well into lower part of the graph, to discourage them from assuming they should be able to take on more enemy aircraft in the following run. If the visual display is too 'heavy' (e.g. if many runs are being debriefed simultaneously), it can instead just be noted for each run whether the random outcomes were above or below the mean, and whether they were within one, two, or more than two standard deviations. Having that information clearly displayed when the relative merits of tactics are being debated can help to qualify whether the apparently 'better' tactic was helped along by chance.

[10] It should be noted that the x-axis is not in units of time, although it is an ordered progression; the number of missiles remaining is monotonically decreasing in time.

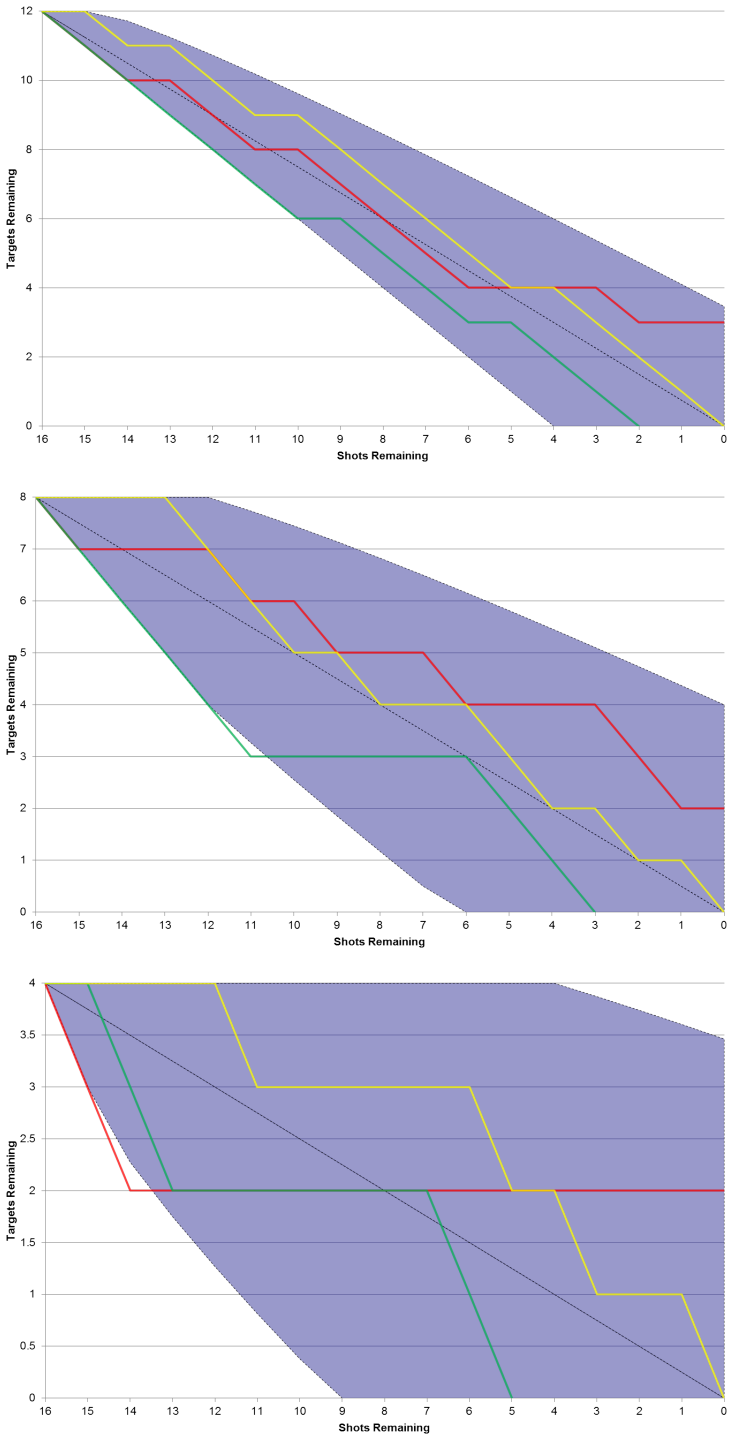

Figure 3: Visual display of 2σ spread for a given number of shots against twelve, eight and four targets, for $P_K = 0.75$, $P_K = 0.5$ and $P_K = 0.25$, respectively.

Shots	Percentage likelihood for given number of targets							
	8	7	6	5	4	3	2	1
16	2.71	7.96	18.97	36.98	59.50	80.29	93.65	99.00
15	1.73	5.66	14.84	31.35	53.87	76.39	91.98	98.66
14	1.03	3.83	11.17	25.85	47.87	71.89	89.90	98.22
13	0.56	2.43	8.02	20.60	41.57	66.74	87.33	97.62
12	0.28	1.43	5.44	15.76	35.12	60.93	84.16	96.83
11	0.12	0.76	3.43	11.46	28.67	54.48	80.29	95.78
10	0.04	0.35	1.97	7.81	22.41	47.44	75.60	94.37
9	0.01	0.13	1.00	4.89	16.57	39.93	69.97	92.49
8	0.00	0.04	0.42	2.73	11.38	32.15	63.29	89.99
7	0.00	0.01	0.13	1.29	7.06	24.36	55.51	86.65
6	0.00	0.00	0.02	0.46	3.76	16.94	46.61	82.20
5	0.00	0.00	0.00	0.10	1.56	10.35	36.72	76.27
4	0.00	0.00	0.00	0.00	0.39	5.08	26.17	68.36
3	0.00	0.00	0.00	0.00	0.00	1.56	15.63	57.81
2	0.00	0.00	0.00	0.00	0.00	0.00	6.25	43.75
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	25.00

(a) $P_K = 25\%$

Shots	Percentage likelihood for given number of targets							
	8	7	6	5	4	3	2	1
16	59.82	77.28	89.49	96.16	98.94	99.79	99.97	100.00
15	50.00	69.64	84.91	94.08	98.24	99.63	99.95	100.00
14	39.53	60.47	78.80	91.02	97.13	99.35	99.91	99.99
13	29.05	50.00	70.95	86.66	95.39	98.88	99.83	99.99
12	19.38	38.72	61.28	80.62	92.70	98.07	99.68	99.98
11	11.33	27.44	50.00	72.56	88.67	96.73	99.41	99.95
10	5.47	17.19	37.70	62.30	82.81	94.53	98.93	99.90
9	1.95	8.98	25.39	50.00	74.61	91.02	98.05	99.80
8	0.39	3.52	14.45	36.33	63.67	85.55	96.48	99.61
7	0.00	0.78	6.25	22.66	50.00	77.34	93.75	99.22
6	0.00	0.00	1.56	10.94	34.38	65.63	89.06	98.44
5	0.00	0.00	0.00	3.13	18.75	50.00	81.25	96.88
4	0.00	0.00	0.00	0.00	6.25	31.25	68.75	93.75
3	0.00	0.00	0.00	0.00	0.00	12.50	50.00	87.50
2	0.00	0.00	0.00	0.00	0.00	0.00	25.00	75.00
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	50.00

(b) $P_K = 50\%$

Shots	Percentage likelihood for given number of targets							
	8	7	6	5	4	3	2	1
16	99.25	99.84	99.97	100.00	100.00	100.00	100.00	100.00
15	98.27	99.58	99.92	99.99	100.00	100.00	100.00	100.00
14	96.17	98.97	99.78	99.97	100.00	100.00	100.00	100.00
13	91.98	97.57	99.44	99.90	99.99	100.00	100.00	100.00
12	84.24	94.56	98.57	99.72	99.96	100.00	100.00	100.00
11	71.33	88.54	96.57	99.24	99.88	99.99	100.00	100.00
10	52.56	77.59	92.19	98.03	99.65	99.96	100.00	100.00
9	30.03	60.07	83.43	95.11	99.00	99.87	99.99	100.00
8	10.01	36.71	67.85	88.62	97.27	99.58	99.96	100.00
7	0.00	13.35	44.49	75.64	92.94	98.71	99.87	99.99
6	0.00	0.00	17.80	53.39	83.06	96.24	99.54	99.98
5	0.00	0.00	0.00	23.73	63.28	89.65	98.44	99.90
4	0.00	0.00	0.00	0.00	31.64	73.83	94.92	99.61
3	0.00	0.00	0.00	0.00	0.00	42.19	84.38	98.44
2	0.00	0.00	0.00	0.00	0.00	0.00	56.25	93.75
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	75.00

(c) $P_K = 75\%$

Table 3: Probability of a given number of targets avoiding more than a given number of shots, given a P_K.

Over the course of the trial, it may also become more clear what the P_I is based on the simulation results. The starting value of the y-axis can then be reduced by this factor, to only include those outcomes that are down to random chance. As a planning tool, the overall P_K is the more appropriate choice.

5.2 Potential Mitigation Strategies

As noted above, small sample sizes are not limited to simulations – realistic air combat encounters are subject to the same laws of statistics. It is therefore not realistic to remove probabilistic elements from simulation. It is also not desirable from a human factors perspective: bored participants or those who can predict the outcome will lead to poor lessons learnt. However, there are a few possible mitigations that can eliminate some of the most insidious effects.

- As an alternative to directly focusing on the outcomes of the encounters (with the statistical variability of the missile outcomes included), one can instead focus on the number of shot opportunities created by a given tactic. One can then later apply the expected distribution of outcomes to the shot opportunities on either side, to generate a statistical view of the probability of success. However, this approach is not without significant challenges. In particular, the outcome of the early missile firings will affect how the encounter develops, creating a branching tree of possible outcomes. It is the author's experience that after perhaps the second or third volley the encounter will necessarily have devolved from any sort of strict plan, due to variations in individual responses and actions. It should also be noted that shot opportunities may be affected by the altitude and speed of each formation, and so the range of each variable for which the results are assumed to e valid should be carefully considered.

- In a similar vein, one can consider fixing the outcomes of the first one or two volleys of shots, and exploring every possible combination. It can be usefully applied in particular to the first shots in the encounter, however note that even two shots will have four possible outcomes, without even considering the outcome of any return fire. It will also be important to ensure that the various runs arn not presented in a known order, to avoid the "fighting the simulation" effect. Conceptually this should not be difficult to implement programmatically, but the author has not seen this feature in a simulator (other than fixing the outcome to 'always' or 'never' successful).

- To mitigate the deceiving effect of regression to the mean (see Section 4.1), runs that will be directly compared in the analysis can be separated in time. This will help to prevent a particularly good or bad run with a particular configuration immediately being compared against a comparator,[11] which

[11] i.e. the same setup except with a different aircraft configuration, weapon, or tactic.

will almost surely appear worse than the good run or better than the bad run, respectively. This will not effect any quantitative results, but is very important qualitatively; it can help prevent the formation of strong biases and unfounded preferences in the participants. However, this strategy can be difficult to implement, as generally the simulator operators will want to group similar runs together for ease of setup, and pilots will prefer to brief to a set of similar missions at the same time.

6 Conclusion

HIL simulation of air combat is an attractive method for examining tactics and comparing aircraft capabilities, but it must be approached with caution. The statistician or tactician who desires a clear, quantitative conclusion will by and large be disappointed; if the answer were that clear, generally simulation would not even be necessary. The goal of the analyst becomes not so much one of decision support, but *indecision* support. However, that is not to say that nothing is learnt through these encounters. There are qualitative tactical lessons to be gained, and quantitative input is important to ensure that random factors are not obscuring those lessons. Additionally, once an effective tactic is found, statistics can be used to quantify the *risk* of failure due to inevitable random outcomes. More specific recommendations based on this general insight will be developed below.

6.1 Recommendations to the Analyst

Given their background, the temptation of the analyst is often to jump to whatever quantitative outputs are available, and start making tables and charts. What is important to recognize is that in the simulator *the probabilities of random events are an input*, and so analysis of outcomes can often turn into a simple validation of inputs. It is the author's experience that pilots are very eager to hear about kill ratios and missiles per target ratios, but as thoroughly documented above these numbers are suspect with the sample size involved. Even if this is well explained, it is human nature that as soon as numbers are put in front of people, they will start to make comparisons and form opinions based upon them – even if they are statistically meaningless. The following recommendations are therefore offered:

- Avoid presenting averages, or at least unannotated averages. Focus instead on noting whether a particular run was within expectations, or was more of an outlier.
- Where possible, focus on the aspects of the encounter that are directly attributable to a chance in tactic – e.g. changes in the number of shot op-

portunities – rather than the overall outcome, with its attendant stochastic variability.

- Highlight runs that were particularly lucky to temper enthusiasm for using them as models of tactical success. Similarly, making sure to point out when a strategy is in danger of being thrown out after one unlucky run. It is the author's experience that the second problem is the more tenacious, especially when participants have prejudged the outcome.

- Pick your battles. The trial is rarely, if ever, designed for your own benefit, and so focus on rectifying the most egregious issues, and let small things go. One also needs to build trust to be listened to.

- Be self-aware of your own biases. The analyst is just as susceptible as the layman to well-known cognitive biases and forming opinions based thereon, and so be careful of reading and presenting numbers in ways that confirm your own bias. Being self-deprecating and using analogies to less charged subjects such as sports to illustrate biases will also help to ensure one does not appear too preachy.

6.2 Recommendations to the Trial Director, Tactician or Requirements Developer

Although much of the discussion above emphasizes uncertainty, this is not to suggest that nothing valuable can be gained from HIL simulation trials. The following recommendations are offered to help focus not just trial planning and analysis, but tactics development for the world of small sample sizes in which modern fighter aircraft operate:

- Match-ups between aircraft should be looked at in terms of probability of success and risk levels, rather than exchange ratios. The assumption is that a war of attrition is not the goal, but rather the protection of valuable assets. This is particularly important when one is at a speed disadvantage to one's opponent – the risk of not knocking one's opponent out of the sky is of paramount importance.

- The robustness of one's tactic to a streak of 'unlucky' missile results will continue to be important, even for quite high P_K. While acquiring a missile with better 'dice' at end-game may help, it is not the whole solution.

- To put a greater emphasis on the risk of negative outcomes, consideration should be given to having at least the initial sequence of 'random' missile outcomes could instead be pre-programmed. Dealing with all possible permutations of missile outcomes will almost certainly be impractical, but in most cases it can be conjectured that early outcomes will have a disproportionate effect. This could help ensure more of the potential

outcome space is covered more efficiently, and also more evenly between different configurations of aircraft, tactic, and threat.

References

[1] Eurofighter Typhoon, "Eurofighter: Swing role," 2012. http://www.eurofighter.com/eurofighter-typhoon/swing-role/mission-configuration/swing-role0.html.

[2] A. Agresti and B. A. Coull, "Approximate is better than "exact" for interval estimation of binomial proportions," *The American Statistician*, vol. 52, pp. 119–126, May 1998.

[3] J. Sauro and J. R. Lewis, "Estimating completion rates from small samples using binomial confidence intervals: Comparisons and recommendations," in *Proceedings of the Human Factors and Ergonomics Society*, 2005.

[4] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, vol. 185, pp. 1124–1131, 1974.

[5] A. Tversky and D. Kahneman, "Belief in the law of small numbers," *Psychological Bulletin*, vol. 76, no. 2, pp. 105–110, 1971.

[6] T. Gilovich, R. Vallone, and A. Tversky, "The hot hand in basketball: On the misperception of random sequences," *Cognitive Pyschology*, vol. 17, pp. 295–314, 1985.

[7] P. Ayton and I. Fischer, "The hot hand and the gambler's fallacy: Two faces of subjective randomness?," *Memory & Cognition*, vol. 32, no. 8, pp. 1369–1378, 2004.

[8] C. J. R. Roney and L. M. Trick, "Sympathetic magic and perceptions of randomness: The hot hand versus the gambler's fallacy," *Thinking & Reasoning*, vol. 15, pp. 197–210, May 2009.

[9] C. Rampell, "The beginning of the end of the census?," *New York Times*, May 2012.

[10] F. Galton, "Regression towards mediocrity in hereditary stature," *The Journal of the Anthropological Institute of Great Britain and Ireland*, vol. 15, pp. 246–263, 1886.

List of symbols, abbreviations, acronyms and initialisms

BVR Beyond Visual Range

HIL Human-In-the-Loop

P_I Probability of Intercept

P_K Probability of Kill

P_L Probability of Launch

PRNG Pseudo-Random Number Generator

P_S Probability of Success

STOVL Short Take-Off Vertical Landing

TTP Tactics, Techniques and Procedures