Cognitive Bias in the Verification and Validation of Space Flight Systems

Steve Larson
Jet Propulsion Laboratory, California Institute of Technology
4800 Oak Grove Dr.
Pasadena, CA 91109
818-354-0679
Steven.A.Larson@jpl.nasa.gov
This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration

Abstract— **Cognitive bias is generally recognized as playing a significant role in virtually all domains of human decision making. Insight into this role is informally built into many of the system engineering practices employed in the aerospace industry. The review process, for example, typically has features that help to counteract the effect of bias. This paper presents a discussion of how commonly recognized biases may affect the verification and validation process.**

Verifying and validating a system is arguably more challenging than development, both technically and cognitively. Whereas there may be a relatively limited number of options available for the design of a particular aspect of a system, there is a virtually unlimited number of potential verification scenarios that may be explored. The probability of any particular scenario occurring in operations is typically very difficult to estimate, which increases reliance on judgment that may be affected by bias. Implementing a verification activity often presents technical challenges that, if they can be overcome at all, often result in a departure from actual flight conditions (e.g., 1-g testing, simulation, time compression, artificial fault injection) that may raise additional questions about the meaningfulness of the results, and create opportunities for the introduction of additional biases. In addition to mitigating the biases it can introduce directly, the verification and validation process must also overcome the cumulative effect of biases introduced during all previous stages of development.

A variety of cognitive biases will be described, with research results for illustration. A handful of case studies will be presented that show how cognitive bias may have affected the verification and validation process on recent JPL flight projects, identify areas of strength and weakness, and identify potential changes or additions to commonly used techniques that could provide a more robust verification and validation of future systems.

TABLE OF CONTENTS

1. INTRODUCTION

The role of cognitive bias in decision making has been the subject of interest in many domains: economists, marketing professionals, and even the intelligence community. But it does not appear to have been systematically employed in the aerospace business.

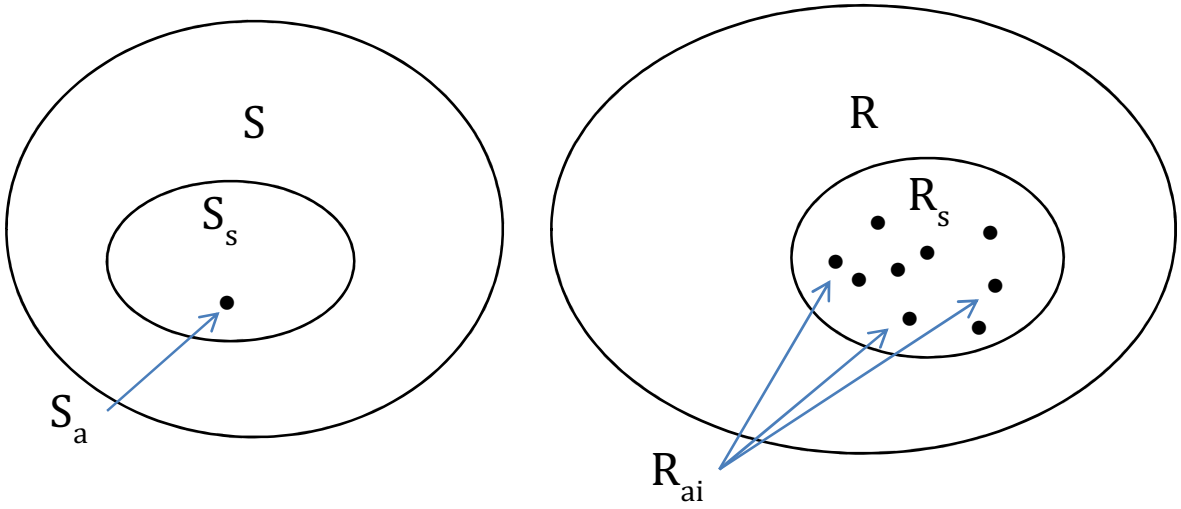
Although the field was recognized in the early 20th century, the touchstone work was published in 1974 by Amos Tversky and Daniel Kahneman [16]. Tversky and Kahneman outlined three biases: Representativeness, Availability, and Anchoring & Adjustment. Subsequent research explored variations on these topics, and opened investigation on other potential bias phenomena such as Confirmation Bias, Overconfidence, and so-called "Illusions of Control".

Cognitive science professionals have attempted to explain bias in terms of a mechanistic model of the mind. Unfortunately, our understanding of the brain does not support anything more than notional models of cognitive function: concepts such as short-term and long-term memory refer to established facts, but we do not understand the actual mechanisms involved. As a consequence, it is not clear whether the dozens of named biases are in fact distinct phenomena, or are instead manifestations of a more general principal rooted in how our minds are organized and operate. Eliminating confounding factors, and understanding the influence of gender, cultural, and other influences remains a challenge for the conduct of research as well as its broader application.

Consequently, the analysis and examples in this paper will be limited to discussion of how bias may affect system verification, and point to potential mitigations. Examples intended to illustrate one bias may also include discussions of other biases, and their possible interactions.

978-1-4577-0557-1/12/$26.00 ©2012 IEEE

Figure 1.



2. The Verification Problem

For purposes of this paper, the problem of system verification will be formulated as follows. Given a set of test results of the actual system (R_{ai}) lying within (or possibly mostly within[1]) the subset of all possible results R that satisfy the system success criteria (R_s), what is the probability that the actual system (S_a) is a member of the subset of possible systems S that represent systems that will always satisfy the success criteria (S_s)?

Even when S is constrained to minor variations on the actual system design, the subset S_s is still quite small compared to S. Likewise, even when R is constrained to results that are meaningfully distinct, R_s is small, and R_{ai} much smaller still. The actual probabilities are generally not feasible to calculate, and even estimates are likely to be worthless.

In practice, systems are accepted based on confidence that they have a qualitatively "high" probability of success, and/or a "low" probability of failure. Test results play an important part in this judgment[2].

3. Cognitive Biases

Dozens of cognitive biases have been identified. Not all are distinct enough to warrant separate treatment. The material to follow presents a handful of well-researched biases that have prospective application to the task of system verification. In each subsection a new bias will be briefly described, followed by an account of salient empirical results, and a discussion of their potential applicability. Case studies will be presented where possible.

The alert reader will no doubt begin seeing commonalities and interrelations between biases. I believe this is entirely warranted. Cognitive science literature allows for common causes underlying seemingly disparate phenomena, and in some cases even suggests them. However, the state of our understanding of the mechanisms of thought and cognition in general is not sufficient to support any but the most notional theory. The material presented here will refrain from speculation on potential underlying causes. The reader, however, is encouraged to consider the ways in which biases may be interrelated, either in a reinforcing or counteracting manner. Some discussion along those lines will be provided in the concluding section.

Representativeness

A representativeness bias is said to exist when a person estimates the likelihood of an outcome based on the similarity between the information available and the outcome itself. Estimates with this bias occur even when subjects are provided with quantitative information on the probabilities involved. The effect is starkly illustrated by an experiment conducted by Tversky and Kahneman [16]. Test subjects were told that a group of individuals consisted of 30 engineers and 70 lawyers. When asked the probability that an individual randomly selected from the group would be an engineer, they correctly responded 30%. However, when provided a description of the person selected that had characteristics of a stereotypical engineer, estimates of the likelihood that the individual was an engineer rose to 50%, despite the fact that the description provided no useful information on what the person actually did for a living.

The preceding is an example of the more general phenomenon of neglecting base rates when estimating probabilities. A common result of base rate experiments is that rare events are typically overestimated, while the likelihood of common events is underestimated. As Koehler, et al, showed, this affects even trained professionals (in this case, doctors) asked to evaluate probabilities in their own domain [7].

[1] It is common to accept systems with small deviations from specified behavior or performance.

[2] Other factors such as the results of probabilistic risk assessments, design characteristics, and so forth, also make important contributions to the overall judgment that a system is ready for deployment.

Tversky and Kahneman explored several other aspects of representativeness. They found that people judged the likelihood of the sequence H-T-H-T-T-H as more likely than the sequence H-H-H-T-T-T to be the result of a series of coin tosses, even though the probabilities of either sequence are the same. Their hypothesis was that the increased estimate of the likelihood of the first sequence was based on the similarity between the sequence itself and the characteristics of the underlying process (i.e., randomness). In a related finding, they found that people have greater confidence in predictions based on a set of results that are internally consistent or redundant, despite the fact that this violates the statistics of correlation—correlated inputs actually decrease the accuracy of predictions.

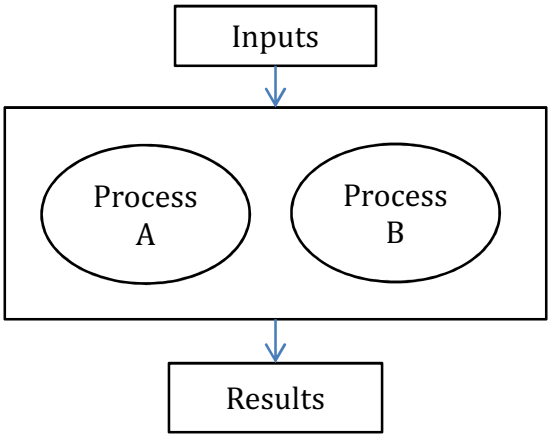


Figure 2. Attributing Results to a Process

To illustrate how representativeness might affect judgments in verification, consider the situation shown in Figure 2. A set of inputs (the test case) is put into a "black box" that contains two processes, A and B. To which process would we attribute the results obtained? If representativeness were operating, those evaluating the results would be more likely to attribute the results to process A, if they were more similar to what they expect process A to produce.

If process A were a system that always operates correctly, while system B does not, there would be a bias towards assuming that a set of correct results[3] was the product of process A. This effect would be reinforced if the test team had already formed an opinion that the system under test was reliable.

The rules of probability would not support this conclusion, at least not to the degree one might suppose. Referring back to the model of the verification problem in section 2, the situation is this: we have only one system under test, but its exact behavior under a variety of conditions is either not known, or not known with certainty. Therefore, it belongs to one of two sets of possible systems; those that fully satisfy requirements, and those that don't. In this example, Process A corresponds to the ensemble of solutions S_s that satisfy requirements under all circumstances, and Process B corresponds to $\overline{S_s}$ = S – S_s, the ensemble of systems that do not satisfy requirements under all circumstances. So, given successful test results R, what is the probability that R is the result of a 100% reliable system, or one that is less reliable?

For notational convenience, we will return to the Process A and B convention. The probability that a successful test outcome is the product of a flawed system is the conditional probability P(B|R). The probability that the same test outcome is the product of a completely reliable system is P(R|A). Using Bayes Theorem, the ratio of the two probabilities is given by:

$$\frac{P(A|R)}{P(B|R)} = \left(\frac{P(R|A)}{P(R|B)}\right) * \left(\frac{P(A)}{P(B)}\right) \quad (1)$$

The first operand on the right corresponds to the ratio of the reliabilities of ensembles A and B, respectively, while the second corresponds to the probability that our system lies within S_s or $\overline{S_s}$. It might appear that that is precisely the ratio our verification seeks to quantify, but it isn't. It's the ratio between the number of systems in S_s and $\overline{S_s}$. Simplification and substitution using the law of total probability we get:

$$P(A|R) = \frac{x\alpha}{1+x\alpha}, \text{ where} \quad (2)$$

$$x = \left(\frac{P(R|A)}{P(R|B)}\right), \text{ and} \quad (3)$$

$$\alpha = \left(\frac{P(A)}{P(B)}\right) \quad (4)$$

There are a few important cases to consider. First, if we presume that design and manufacturing quality is such that the likelihood that an out-of-spec error has been introduced is 50%, then α is unity. Suppose the potential errors we have allowed for will still result in a system that is 99% reliable. The probability that a successful test result indicates that our system is indeed a member of the set of 100% reliable systems would only be 50.25%. This is hardly better than our going-in assumption of how well we had constrained the solution space we needed to consider. In other words, for a design and manufacturing process that is well-constrained enough to ensure that all systems that are produced by it have very high reliability, a successful test program cannot tell us much beyond our *a priori* assumptions about how likely it is that the process will produce an unreliable system.

In the more likely scenario where $\overline{S_s}$ is much larger than S_s, α is rather small. If the ratio of reliabilities of the resulting ensembles is still close to unity, then for very small α equation (2) simplifies to P(A|R) = α. In other words, if we acknowledge that our process controls will allow for a

[3] Correct test results are the only case of interest—systems are not generally deployed with failures in the test program unless they are waived or otherwise accepted at "good enough".

relatively large set of possible systems that are very good, but not quite perfect, then positive test results give us no real information on whether we truly have built the reliable system we set out to.

Most complex systems are currently verified at least in part via the use of simulations, since in many cases it is either impossible, or unacceptably dangerous, to test all conditions using a live system. This further muddies the waters. Our ability to infer total system reliability from a combination of simulated and real component performance data depends entirely on the fidelity of the simulation. Unless this fidelity is anything other than 100% (at least in the features that affect testing) then it will provide even less information than the already disappointing results just described.

If the population of possible systems that are close, but not quite close enough, to the desired level of reliability is larger than the population of reliable ones, then a given set of tests will comprise a smaller sampling of that population than it would of the satisfactory system population. Thus, a test set that adequately samples the population of closely related satisfactory systems, will likely undersample the population of unsatisfactory systems, with concomitant implications for the reliability of conclusions drawn on the basis of their results. The fact that people, even those trained in their use, tend to be insensitive to the statistics of sample size implies that this disproportion further exacerbates bias due to representativeness.

Availability

The availability heuristic posits that people rate uncertain events as more or less likely according to how easily they can recall instances of the event having occurred in the past, or how easily they can construct a (subjectively) credible scenario where the event would occur. Studies have shown that this works in reverse: uncertain events are estimated as being less likely if it is difficult to recall or imagine scenarios where the event has, or could, occur [15].

This tendency has important implications in both development and verification. If developers of a system have difficulty recalling or imagining an adverse scenario ("that's never happened before", or "I can't see how that's possible") they are unlikely to invest significant effort in designing defenses against an event perceived as unlikely. If this conclusion is conveyed to the verification team, other heuristics may come into play: familiarity and anchoring, for example. Engineers tasked with verifying the system may likewise assess the probability of the event as being (again, subjectively) improbable, and thus may even conclude that no test is necessary. Indeed, if a verification team cannot construct a scenario where an adverse event would occur, they cannot construct a test, or if they do, it will most likely be ineffective.

NASA's Phoenix mission illustrates the difficulty, even when a general failure scenario can be constructed. In order to save power and transmit all of its acquired data to Earth, the lander would go into a sleep mode during the Martian night, and wake up periodically to transmit data to relay satellites passing overhead. Alert operators discovered that it was taking longer and longer for the computer to reboot when exiting sleep mode. If the trend continued too long, reboots would take longer than the hard-wired watchdog timer setting, which would effectively end the mission. Investigation revealed that a rare combination of conditions would trigger a chain of events that would eventually lead to a lengthening of the reboot time. Engineers correctly identified sleep/wake cycling as a critical function, and it was thoroughly tested. However, the precise combination of events was not foreseen and built into the tests. Even if engineers had identified the circumstances as a potential test case, the chain of reasoning leading from cause to ultimate consequence was sufficiently arduous as to make it unlikely that it would have been singled out as an important case among the large number of possible cases.

Sherman, et al [15] were also able to show that the relative concreteness of language used to describe a potential failure has a significant influence on others' estimates of likelihood. Causes of an event that are described in vague terms increase the difficulty encountered in imagining a failure scenario, while specific language increases estimates of likelihood. This has important implications for training. Engineers must be trained to be as specific and concrete as possible when raising concerns about potential failure modes, and conversely, they must be trained to take vaguely expressed concerns seriously enough to explore them in more concrete terms instead of dismissing them out of hand.

Anchoring and Adjustment

The anchoring and adjustment heuristic was described by Tversky and Kahneman in 1974, and subsequent research has elaborated upon it. In brief, it has been observed that people make estimates based on an initial answer, and make adjustments from there based on additional information. As Tversky and Kahneman put it, "adjustments are typically insufficient." [16]

Although the phenomenon was initially described in terms of anchoring on initial values when arriving at numeric estimates, the results have been found to hold in more general cases. One arena in engineering where this effect can be seen is in the use of so-called "heritage". Whether at the component, system, or concept level, the initial solution people have in mind when they first consider a problem can dominate the entire development process. Of course, use of proven ideas, designs, and components is a valuable tool in reducing cost and development time, as well as improving reliability.

The Mars Climate Orbiter may provide an example of how anchoring on a known quantity might adversely affect reasoning. In a 1999 report, the Mishap Investigation Board found, "the operations navigation team supporting MCO to be somewhat isolated from the MCO development and operations teams, as well as from its own line organization,

by inadequate communication. One contributing factor to this lack of communication may have been the operations navigation team's assumption that MCO had Mars Global Surveyor (MGS) heritage and the resulting expectation that much of the MCO hardware and software was similar to that on MGS." [11]

Overconfidence

Stuart Oskamp [12] conducted a study correlating the relationship between level of confidence and level of accuracy. The results have important implications for system verification.

In the study, a group of trained experts were provided with progressively more information about a case, and then asked to answer a set of multiple choice questions based on their conclusions. Their answers, overall accuracy, and self-assessment of their confidence were recorded after each of four increments of information was provided.

The difference between the experts' accuracy (~28%) and random chance (20%) was nonsignificant. Moreover, there was no significant change in accuracy with successive increments of information. However, confidence in their answers increased monotonically from approximately 33% to 53% over the course of the experiment. As the test progressed, respondents made fewer changes to their previous answers, reflecting their increased confidence.

One implication of the Oskamp study is that more testing may not necessarily be better, at least from the perspective of drawing accurate conclusions about system readiness. It would be foolish to deploy a system without verifying that it met its specifications. Requirements testing is a reasonably well-bounded problem, but where should we draw the line when it comes to testing off-nominal conditions? This is a potentially infinite space, and yet it is the "outlier" cases that often bring systems down. Although I will have more to say about this later, it seems warranted at this point to emphasize that quality of testing is probably more important than quantity.

One might object that studies done in the context of clinical psychology might not be applicable to engineering. While it is always reasonable to question extrapolation of results from one domain to another, overconfidence has been the subject of many studies, including those examining experts in technical domains. In a 2004 survey of the state of the practice, Ulrich Hoffrage reported that overconfidence was the dominant finding across all studies [5]. Research findings on the calibration of experts within their domain of expertise varies from nearly perfect in the case of weather forecasters[4], to little better than chance in one study of lawyers predicting case outcomes.

It is tempting to believe that, as engineers, our performance would be similar to the weather forecasters, who operate in the realm of theoretical models and measurement. However, Wilson *et al* observed that people are notoriously bad at recognizing bias in themselves, and equally poor in correcting their own bias even when they recognize it [18]. In a survey on debiasing Baruch Fischhoff found, "particularly striking…the lack of differences in…experts making judgments in their own fields." It would appear that overconfidence should be assumed until proven otherwise.

Indeed, overconfidence is often assumed in the review of aerospace projects. We expect, among other things, that reviewers will recognize, and hopefully correct, overly optimistic assessments of system readiness. Unfortunately, we may be overconfident in their effectiveness. Replacing, or at least augmenting the judgment of, a single individual (e.g., a project manager) with a review board is effectively acknowledging that there is nothing to be done about the bias presumed in the individual subject their scrutiny. While this may ameliorate biases affected by motivational factors, it isn't necessarily the most effective strategy. Training seems to be the most effective debiasing strategy [2], at least as far as confirmation bias goes. Unfortunately, this may require more time and data relevant to the individuals involved[5] than is practical. In very large projects, review boards charged with assessing mission readiness generally do not have access to all the relevant information that would allow them to home in on specific areas where overconfidence might be a problem. Time constraints may also limit their ability to dig deep enough to identify problem areas. Persistent controversy over the actual versus claimed capabilities of large military projects such as ballistic missile defense systems serve as a case in point.

Illusions of Control

Ellen Langer [8] defined an illusion of control as "an expectancy of a personal success probability inappropriately higher than the objective probability would warrant." In a classic experiment Langer found that allowing participants to select a specific lottery ticket dramatically increased their belief that they held the winning ticket. The same study also found that participants who were familiar with the images printed on the tickets (football players) were more optimistic about their chances of winning than those who were less familiar.

These results were later generalized into a group of factors relating to skill, where aspects of a situation involving knowledge, choice, competition, or active engagement act as cues that mislead participants into believing they have control where little or none exists. For example, Alloy and Abramson conducted an experiment where participants were asked to press a button to cause a light to turn on. The light was programmed to turn on at one of two fixed rates (25% or 75%) regardless of how the button was pressed.

[4] This does not mean that weather forecasters were found to be perfectly accurate, merely that their confidence in their accuracy was in near perfect correspondence with their actual ability to predict the weather.

[5] The most effective training uses feedback on the trainee's own confidence versus actual accuracy.

Nevertheless, participants all believed they had some control over the operation of the light, with estimates of control being higher when the light came on 75% of the time, and lower in the 25% case. It appears that the mere act of engaging in the task created the illusion of control.

Other studies have found that previous success influences people to believe they can predict chance events such as a coin toss [9], and that the stress relief accompanying a belief in control enhances the illusion [3].

System verification contains elements of both skill and chance. The skill aspect hardly needs explanation. Test engineers bring to bear considerable knowledge of the system, its planned use, and related knowledge of materials and the like. But as the research has shown, the opportunity to apply skill to a task that also contains elements of chance can lead to an exaggerated sense of control over the outcome.

The chance element enters into the picture when we consider the fact that in the "real world" the deployed system will interact with users and the environment, both of which are rich sources of stochastic input. If we define success in verification as showing that the system will perform as expected over its lifetime, then successful verification is unquestionably a matter of both skill and luck. There is simply no way to test more than a tiny fraction of all possible scenarios. Furthermore, since the number of potentially dangerous latent flaws in the system is itself unknown, there is no way to know how much additional confidence one should have in the system, given that it has passed any particular test or tests.

Confirmation

Science and engineering professionals are quite familiar with the notion of confirmation bias. It is standard practice to assign the task of verification to people who were not involved in the development to avoid tainting the results. But research in cognitive science has shown that there is more to confirmation bias than we might suspect.

We begin with a definition: "'Confirmation bias' means that information is searched for, interpreted, and remembered in such a way that it systematically impedes the possibility that the hypothesis could be rejected." [13] Although this definition would include the practice of seeking, or accepting, only information that confirms one's previous belief, it is clear that confirmation bias involves much more that. It includes effects produced by the way we retrieve and store information as memories, and the strategies we use to test theories.

In a classic early experiment on the subject, Peter Wason conducted a simple experiment involving a sequence of numbers. Participants were given the sequence "2, 4, 6", and asked to come up with other three-number sequences to determine the underlying rule. For each guess they would be told whether or not their series fit the rule. When they had gathered enough data to be confident they had correctly determined the rule, the "testing" stopped, and they were told whether they had determined the rule correctly or not. [17]

The results were revealing. Most participants guessed a rule (e.g., even numbers, multiples of the series "1, 2, 3", etc) and tested their theories by coming up with examples that fit their rule. Participants were allowed as many trials (a yes or no answer to whether a proposed series matched the rule) as they wanted before stating what they believed the rule to be. Most (~80%) required more than one round of trials before correctly identifying the rule. The rule itself was simple: any series of increasing numbers. Wason found that the people who had the most difficulty failed to test series that would falsify their hypothesis, whereas those who did needed fewer rounds of trials.

The strategy employed by most participants is known as a positive test strategy (PTS). A PTS has the advantage of simplicity and efficiency, as it is relatively straightforward to generate cases that support a hypothesis, and it requires relatively fewer trials. It is also an important part of any system verification. The assertion that the system meets a requirement becomes the hypothesis, and the verification activities performed provide the evidence to support or falsify the hypothesis. A PTS *per se* does not introduce a confirmation bias unless the results comprise a subset of the correct hypothesis (i.e., the actual system behavior). The problem is that unless the test program includes all cases that would confirm correct behavior, it omits cases that might falsify the notion that the system meets requirements. The following example illustrates how easily a confirmation bias can be introduced into a test program:

> A spacecraft has a solid state recorder that implements a circular buffer, which is very large compared to the units of data it stores. When the C&DH is done sending data it sends a message to mark the end of recording, and in return receives a pointer to the next write location. The system is required to handle both compressed and uncompressed data. The verification program tested all required behavior. However, in operation it was discovered that the unusual combination of using compression and wrapping around the end of the circular buffer caused the recorder to erroneously return an error code. This in turn caused the next write location pointer to not get updated correctly on the C&DH side. Subsequent recording overwrote the previously recorded data.

In general, research has not found that people deliberately seek only that information that confirms their beliefs, or select strategies that are biased towards a particular outcome. However, it has confirmed that motivational factors can influence the way information is retrieved or evaluated. Sanitoso *et al* found that individuals recalled more instances of events that confirmed their possession of positive traits, and that these memories were recalled more

quickly [14]. Lord et al found that participants were more critical of evidence that refuted their beliefs than information in support [10]. Other researchers [4] have found that people give greater weight to evidence that supports a previously adopted hypothesis.

These findings suggest that the usual strategy of populating a test team with individuals who were not involved in development may not be enough. If test engineers have any personal stake in the "goodness" of the system under test (e.g., company pride, personal regard for the developers, or personal bias towards an aspect of the technical solution), their selection of test cases and interpretation of test data may be subtly biased.

Mitigating this problem can be challenging. Hiring a separate company to test a system does not guarantee that other biases are not thereby introduced. However, measures such as surveying team members for their attitudes towards the product and development team may alert leadership to potential problems. These may be ameliorated through judicious selection of reviewers who can be called in at various points in the process

4. DEBIASING

The biases and heuristics research program of the 1970's and 1980's eventually stimulated work challenging its conclusions, or at least bounding the conditions under which bias occurred. Efforts to explain away findings of bias as artifacts of flawed methodology have not been particularly successful, nor have efforts to eliminate bias. The latter result is a disappointment to those concerned with its potentially disastrous effects (e.g., a decision to operate a manned mission that is not ready). In this section I will outline the findings in two reviews that examined the debiasing problem.

Baruch Fischhoff reviewed the literature specifically on the hindsight and overconfidence biases [2]. He presented a framework of categories for debiasing strategies which centered on either modifications to the task presented to individual judges, the judges themselves, or a combination of the two. Although the intent of the framework was related more towards characterizing the research, it provides a useful reference for discussing strategies for countering bias in "real life".

It might first appear that we have little choice in the problems we face in system verification. And yet Fischhoff identified flaws in research methods that have analogs in system verification. First off, clarity in the verification task is important. Testers should know exactly what questions they are expected to answer. Vague directives like, "just tell us if it's going to work" are not helpful. Test engineers may also be sensitive to what may be termed the "political" environment. If the perceived goal of the test program (e.g., "convince the customer we're good to go so we can get stop spending money on testing") is other than the explicitly stated one, then the results may be tainted. Test engineers should also be allowed enough latitude to express their findings in the way that best enables them to communicate, rather than be forced to present their findings in a preset manner. Finally, limit the number and scope of questions put to them. Too much redundancy or irrelevance can lead to knee-jerk responses whose functional purpose is simply to get through the exercise as quickly as possible.

Doubts about the ability of individuals to counter their own biases notwithstanding, Fischhoff identified some approaches to mitigating overconfidence that have met with some success. Alerting people to the possibility of bias is a first step. Identifying the direction and magnitude of the bias provides additional, potentially useful, information. The most effective measures incorporate personalized feedback on the trainee's measurable bias through specific examples and exercises, though care must be taken to do so without creating defensive resistance. Forcing people to explicitly express tacit knowledge, and encouraging the search for falsifying evidence have also been tried.

Some have suggested strategies for compensating for bias by eliminating the biased individual altogether. Proposed alternatives include using a decision-making instrument (e.g., a weighted matrix of desired & undesired characteristics), replacing individuals with teams of experts, using calibration data to correct biases *a posteriori*, or simply acknowledging the existence of bias and using additional or other criteria to make important decisions. Decision matrices are common elements of decision processes, though as anyone who has used one knows, they can be "gamed" easily, and their use does not eliminate bias without careful oversight. Use of teams of experts—in aerospace we often call them review boards—has been discussed previously. Data suitable to correct known biases is generally not available, nor would it be easily obtained. It may be that simply acknowledging the existence of bias and incorporating that awareness into the overall decision-making process may be one of the most pragmatic approaches available.

Wilson *et al* looked at the debiasing problem from a very different angle: mental "contamination" [18]. The theoretical and experimental work they discuss are concerned with efforts at an individual level to detect and correct the influence of bias. Unlike Fischhoff, Wilson *et al* took on debiasing in the general sense, so unless otherwise indicated the following discussion may be applied to any bias.

They present a conceptually simple test to determine whether bias will occur. If the answer to any of the following questions (predicated on the fact that a potentially biased thought process has occurred) is "no", mental contamination is assumed. All "yes" answers indicate the absence of contamination.

1) Is the person aware that a potentially biased thought process has occurred?
2) Are they motivated to correct it?

3) Do they know the direction and magnitude of the bias?
4) Do they have the internal ability to correct it?

This appears to offer a straightforward means to identify the existence of bias (or proof of its absence), as well as hints at a strategy to eliminate it. Unfortunately, as the authors go on to show, this is not the case. Nevertheless, it remains a useful conceptual reference.

Their criteria for establishing the presence or absence of bias are similar to the causes they discuss:

1) Lack of awareness of one's own mental process
2) Lack of control over one's mental process
3) Poor understanding of the existence and nature of biases they may have
4) Inadequate motivation to correct their own bias

The third cause is clearly amenable to education, but mitigating the others may be difficult, if not impossible. It is at least theoretically possible to provide sufficient motivation for people to try to correct their own biases, but doing so would depend on correctly identifying potentially conflicting motivations and the ability to offer appropriate incentives. Causes one and two present the greatest difficulty. Beyond providing some relatively superficial awareness training and cognitive/behavioral modification, addressing these issues is more than a typical organization is willing to contemplate.

Relying on individuals to correct their own bias, whether or not they have had training intended to improve their ability to do so, is not only flawed, but may actually exacerbate the problem. Wilson et al concluded that, "even in the rare instances in which people believe that their judgments are biased, they may not successfully debias these judgments. In fact, their corrected judgments might be worse than their uncorrected ones."

Part of the problem is that people generally have inaccurate theories about debiasing. Prior to being exposed to potentially contaminating information[6], people who anticipate the possibility may take steps to eliminate or reduce the effect. The simplest and most effective by far, means of eliminating bias is via exposure control: it is not possible to be influenced by information that one has not been exposed to. People may also make mental preparations, such as developing counteracting arguments, for dealing with biasing information. Once exposure to biasing information has occurred, people may attempt to resist its influence, correct for it, or simply decide to override their own conclusions when it comes time to act on their beliefs. Interventions at the outset of the process (i.e., before exposure to contaminating information) are far more effective that interventions that occur later in the process. Unfortunately, people tend to have greater faith in their ability to intervene in later stages. This may be due to an illusion of control, following the Cartesian concept that beliefs are accepted after due consideration, when the reality may be that we operate in the reverse, believing everything at first, and then only rejecting those things that we decide to "unbelieve" later.

[6] Here we are concerned with biases relating to the undue influence of information, such as anchoring, or potentially motivating factors. Biases due to the counterintuitive nature of statistical reasoning are excluded.

5. MARS POLAR LANDER ANALYSIS

Following the loss of the Mars Polar Lander (MPL) mission, a special review board was convened at JPL. The board performed a thorough investigation into the incident, and identified with near certainty the proximal cause of the failure. In addition to elucidating the exact nature of the software logic error that caused the descent engines to terminate prematurely, the final report [6] identified a number of contributing factors, including weaknesses in the verification program. This section will analyze the findings contained in the board's final report and discuss them in light of the cognitive biases discussed in this paper.

The fact that the touchdown sensors could produce spurious positive signals was known to the team, and a requirement was written to ignore the inputs prior to reaching 40m altitude. However, the report found that, "the requirement did not specifically describe [events leading to spurious signals], and consequently, the software designers did not properly account for them." This finding points directly at the availability bias, which would lead the test team (and, indeed, the independent Mission Safety and Success Team assembled at JPL during development) to mistakenly assume that nothing was amiss on the basis of their inability to envision scenarios where the logic would fail.

The argument that availability played a role in the MPL failure is bolstered implicitly by the board's recommendation to perform fault tree analysis prior to test planning to "define test cases that are needed to drive out logic paths that must be tested." This recommendation, if followed, would have increased the likelihood that scenarios where the logic failure was possible would have been more easily identified (i.e., "available", in the terms of cognitive science).

The problem was compounded by the fact that the touchdown sensing logic was not tested in flight configuration. As discussed in the section on representativeness bias, this introduced a confounding factor in the reasoning that allowed the project to assume that test results in the simulated environment were directly transferrable to the flight configuration.

The lack of clear information in the requirements may also have contributed an inadvertent confirmation bias, induced by misinterpretation of a positive test strategy. The requirements as written were verified, but because the actual system could behave in ways that were contrary to expectations under certain (untested) conditions, the results were a subset of a larger set of behaviors that included

undesirable ones. These are precisely the conditions that induce a confirmation bias. Indeed, the recommendation that "test teams need to assume that there is an error[and] must examine every requirement on the software to test whether they can identify a set of conditions that could 'break' the software" is precisely the sort of negative test that Tversky and Kahneman identified as a key ingredient to effective testing.

It is important to note that MPL was subjected to an extra level of scrutiny compared to other projects. This was a result of the findings in the investigation into the loss of the Mars Climate Orbiter (MCO), which had occurred only 3 months prior to MPL's scheduled arrival at Mars. The Mission Safety and Success Team mentioned earlier was convened with the purpose of ensuring that problems on MCO did not recur on MPL. Appointing a team of independent experts is one strategy to counteract the assumed bias in those it is charged with overseeing, as discussed in the section on overconfidence. The fact that this outside board looked at much of the same material as the developers and testers, and found no problems either with the design or test program highlights the difficulty of removing bias. Superficially, it seems like a tactic that would likely find errors and biases in the project team's work, but unless great care is exercised in the selection and conduct of the team's activities, it is quite possible that they will fall into the same traps as the project team itself.

MPL was lost due to a vulnerability in the touchdown logic. However, the board found that it may well have encountered problems of equal gravity had the landing been successful. A complex interaction between software logic, parameter settings, and the loss of the receiver during entry, decent, and landing (EDL), could result in the lander never switching over to the backup receiver, and thus never being able to receive commands from the ground again. This dire consequence would have required that the lander also experience circumstances that would trigger a safe mode entry before a backup sequence was able to perform the receiver swap.

The complex logic and somewhat obscure combination of events needed to trigger the problem is another case where availability could be expected to come into play. A review of the test program revealed that the specific circumstances were never tested, and in some instances the parameters involved were not tested in any scenario. Additionally, tests of the post-touchdown functionality assumed commandability (this would not have been the case in the failure scenario), and verified requirements from that point forward. The positive test strategy employed by the project likely contributed to an overconfidence that the system would perform correctly. The problem of availability was also present, and could have been mitigated had the team followed the board recommendation to use "flow diagrams and logic charts [that can be] used to identify test cases that must be run to verify that the logic provides the desired actions."

6. CONCLUSION

Some level of bias is inevitable in system verification. Although cognitive science does not provide many proven techniques to eliminate the problem, it does suggest some general strategies that may mitigate it.

It would appear that availability is perhaps the most significant problem, as most engineering professionals are already aware. Overconfidence due to a variety causes is also ubiquitous. Use of many of the tools and techniques listed below are already common in aerospace systems development and verification, but the results presented here emphasize their importance, and can sharpen our focus in their use:

- Employ "Red Teams" to devise tests designed to find and exploit vulnerabilities in the system design
- Use randomized testing, such as Monte Carlo simulations, to increase test volume and remove the human factor from test case selection
- Develop a culture of tolerance that allows individuals to propose seemingly absurd failure scenarios, and make sure they get a fair hearing
- Provide training to reviewers and testers on the nature and effects of biases that may affect their work
- Begin development of test scenarios early, and revisit and refine then throughout the development process, incorporating changes to system design and new ideas about failure modes
- Avoid speculating on the likelihood of failure scenarios, and challenge assumptions about system behavior and operations
- Use care in the selection of test engineers and reviewers to minimize the potential for bias, and monitor their efforts
- Do not rely on any single mitigation to counteract potential bias
- Employ hazard analysis techniques such as STAMP or HAZOP that shift the focus away from familiar component-level failure analysis and challenge engineers to think of how failures could occur from a different perspective
- Improve development and manufacturing processes to eliminate defects up front rather than hope to discover and fix them in test

REFERENCES

[1] Alloy, L.B. & Abramson, L.Y. (1979) *Judgment of contingency in depressed and nondepressed students: Sadder but wiser?* Journal of Experimental Psychology: General, 108, 441-485

[2] Fischhoff, B. *Debiasing*, in Judgment under Uncertainty: Heuristics and Biases (Kahneman, D., Slovic, P. & Tversky, A. eds.) 1982 Cambridge University Press, Cambridge, UK

[3] Friedland, N., Kienan, G. & Regev, Y. (1992). *Controlling the uncontrollable: Effects of stress on perceptions of controllability*. Journal of Personality and Social Psychology, 63, 311-328.

[4] Gadenne, V. & Oswald, M. (1986) *Entstung und Veränderung von Bestätigungstendenzem beim Testen von Hypothesen* [Formation and alteration of confirmatory tendencies during testing of hypotheses]. Zeitschrift für Experimentelle und Angewandte Psychologie, 22, 360-374.

[5] Hoffrage, U. *Overconfidence*, in Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgment, and Memory. (Pohl, R. ed.) 2004 Psychology Press, New York

[6] JPL Special Review Board, *JPL D-18709 Report on the Loss of the Mars Polar Lander and Deep Space 2 Missions,* 22 March 2000, California Institute of Technology

[7] Koehler, D. Brenner, L. & Griffin, D. (2002) *The Calibration of Expert Judgment: Heuristics and Biases Beyond the Laboratory*, in Heuristics and Biases: The Psychology of Intuitive Judgment (Gilovich, et al , eds.) 2002, Cambridge University Press, Cambridge.

[8] Langer, E. J. (1975), *The Illusion of Control*, Journal of Personality and Social Psychology **32** (2): 311–328

[9] Langer, E.J. & Roth, J. (1975). *Heads I win, tails it's chance: The illusion of control as a function of the sequence of outcomes in a purely chance task*. Journal of Personality and Social Psychology, 32, 951-955.

[10] Lord, C.G., Ross, L. & Lepper, M.R. (1979). *Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence*. Journal of Personality and Social Psychology, 37, 2098-2109.

[11] NASA Office of Chief Engineer, *Mars Climate Orbiter Mishap Investigation Board Phase 1 Report*, 10 November, 1999, National Aeronautics and Space Administration.

[12] Oskamp, S. *Overconfidence in case-study judgments*, in Judgment under Uncertainty: Heuristics and Biases (Kahneman, D., Slovic, P. & Tversky, A. eds.) 1982 Cambridge University Press, Cambridge, UK

[13] Oswald, M. & Grosjean, S. (2004) *Confirmation bias*, in Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement, and Memory, Pohl, R. ed. Psychology Press, New York, 2004

[14] Sanitoso, R., Kunda, Z. & Fong, G.T. (1990) *Motivated recruitment of autobiographical memories*. Journal of Personality and Social Psychology, 59, 229-241

[15] Sherman, S., Cialdini, R., Schwartman, D. & Reynolds, K, *Imagining Can Heighten of Lower the Percieved Likelihood of Contracting a Disease: The Mediating Effect of Ease of Imagery*, in Heuristics and Biases: The Psychology of Intuitive Judgment (Gilovich, T, Griffin, D & Kahneman, D. eds.) 2002, Cambridge University Press, Cambridge, UK

[16] Tversky, A. & Kahneman, D. (1974), *Judgment under Uncertainty: Heuristics and Biases,* Science, 185, 1124-31

[17] Wason, P.C. (1960) *On the failure to eliminate hypotheses in a conceptual task*. Quarterly Journal of Experimental Psychology, 20, 273-281.

[18] Wilson, T. D., Centerbar, D. B. & Brekke, N. *Mental Contamination and the Debiasing Problem* in Heuristics and Biases: The Psychology of Intuitive Judgment (Gilovich, T., Griffin, D. & Kahneman, D. eds.) 2002, Cambridge University Press, Cambridge, UK

BIOGRAPHY

Steve Larson *received an M.S. in Physics from California State University, Northridge in 1989. He has been with JPL for more than 20 years. He worked on the development of a number of Earth orbiting and deep space missions, including the Advanced Spaceborne Thermal Emission and Reflection Radiometer, Thermal Emission Spectrometer, Europa Orbiter, Space Interferometry Mission, and the Gravity Recovery and Interior Laboratory.*