Appendix F
Personal observations on the reliability of the Shuttle

by R. P. Feynman

Introduction

It appears that there are enormous differences of opinion as to the probability of a failure with loss of vehicle and of human life. The estimates range from roughly 1 in 100 to 1 in 100,000. The higher figures come from the working engineers, and the very low figures from management. What are the causes and consequences of this lack of agreement? Since 1 part in 100,000 would imply that one could put a Shuttle up each day for 300 years expecting to lose only one, we could properly ask "What is the cause of management's fantastic faith in the machinery?"

We have also found that certification criteria used in Flight Readiness Reviews often develop a gradually decreasing strictness. The argument that the same risk was flown before without failure is often accepted as an argument for the safety of accepting it again. Because of this, obvious weaknesses are accepted again and again, sometimes without a sufficiently serious attempt to remedy them, or to delay a flight because of their continued presence.

There are several sources of information. There are published criteria for certification, including a history of modifications in the form of waivers and deviations. In addition, the records of the Flight Readiness Reviews for each flight document the arguments used to accept the risks of the flight. Information was obtained from the direct testimony and the reports of the range safety officer, Louis J. Ullian, with respect to the history of success of solid fuel rockets. There was a further study by him (as chairman of the launch abort safety panel (LASP)) in an attempt to determine the risks involved in possible accidents leading to radioactive contamination from attempting to fly a plutonium power supply (RTG) for future planetary missions. The NASA study of the same question is also available. For the History of the Space Shuttle Main Engines, interviews with management and engineers at Marshall, and informal interviews with engineers at Rocketdyne, were made. An independent (Cal Tech) mechanical engineer who consulted for NASA about engines was also interviewed informally. A visit to Johnson was made to gather information on the reliability of the avionics (computers, sensors, and effectors). Finally there is a report "A Review of Certification Practices, Potentially Applicable to Man-rated Reusable Rocket Engines," prepared at the Jet Propulsion Laboratory by N. Moore, et al., in February, 1986, for NASA Headquarters, Office of Space Flight. It deals with the methods used by the FAA and the military to certify their gas turbine and rocket engines. These authors were also interviewed informally.

Solid Rockets (SRB)

An estimate of the reliability of solid rockets was made by the range safety officer, by studying the experience of all previous rocket flights. Out of a total of nearly 2,900 flights, 121 failed (1 in 25). This includes, however, what may be called, early errors, rockets flown for the first few times in which design errors are discovered and fixed. A more reasonable figure for the mature rockets might be 1 in 50. With special care in the selection of parts and in inspection, a figure of below 1 in 100 might be achieved but 1 in 1,000 is probably not attainable with today's technology. (Since there are two rockets on the Shuttle, these

rocket failure rates must be doubled to get Shuttle failure rates from Solid Rocket Booster failure.)

NASA officials argue that the figure is much lower. They point out that these figures are for unmanned rockets but since the Shuttle is a manned vehicle "the probability of mission success is necessarily very close to 1.0." It is not very clear what this phrase means. Does it mean it is close to 1 or that it ought to be close to 1? They go on to explain "Historically this extremely high degree of mission success has given rise to a difference in philosophy between manned space flight programs and unmanned programs; i.e., numerical probability usage versus engineering judgment." (These quotations are from "Space Shuttle Data for Planetary Mission RTG Safety Analysis," Pages 3-1, 3-1, February 15, 1985, NASA, JSC.) It is true that if the probability of failure was as low as 1 in 100,000 it would take an inordinate number of tests to determine it (you would get nothing but a string of perfect flights from which no precise figure, other than that the probability is likely less than the number of such flights in the string so far). But, if the real probability is not so small, flights would show troubles, near failures, and possible actual failures with a reasonable number of trials. and standard statistical methods could give a reasonable estimate. In fact, previous NASA experience had shown, on occasion, just such difficulties, near accidents, and accidents, all giving warning that the probability of flight failure was not so very small. The inconsistency of the argument not to determine reliability through historical experience, as the range safety officer did, is that NASA also appeals to history, beginning "Historically this high degree of mission success..."

Finally, if we are to replace standard numerical probability usage with engineering judgment, why do we find such an enormous disparity between the management estimate and the judgment of the engineers? It would appear that, for whatever purpose, be it for internal or external consumption, the management of NASA exaggerates the reliability of its product, to the point of fantasy.

The history of the certification and Flight Readiness Reviews will not be repeated here. (See other part of Commission reports.) The phenomenon of accepting for flight, seals that had shown erosion and blow-by in previous flights, is very clear. The Challenger flight is an excellent example. There are several references to flights that had gone before. The acceptance and success of these flights is taken as evidence of safety. But erosion and blow-by are not what the design expected. They are warnings that something is wrong. The equipment is not operating as expected, and therefore there is a danger that it can operate with even wider deviations in this unexpected and not thoroughly understood way. The fact that this danger did not lead to a catastrophe before is no guarantee that it will not the next time, unless it is completely understood. When playing Russian roulette the fact that the first shot got off safely is little comfort for the next. The origin and consequences of the erosion and blow-by were not understood. They did not occur equally on all flights and all joints; sometimes more, and sometimes less. Why not sometime, when whatever conditions determined it were right, still more leading to catastrophe?

In spite of these variations from case to case, officials behaved as if they understood it, giving apparently logical arguments to each other often depending on the "success" of previous flights. For example. in determining if flight 51-L was safe to fly in the face of ring erosion in flight 51-C, it was noted that the erosion depth was only one-third of the radius. It had been noted in an experiment cutting the ring that cutting it as deep as one radius was necessary before the ring failed. Instead of being very concerned that variations of poorly understood conditions might reasonably create a deeper erosion this time, it was asserted, there was "a safety factor of three." This is a strange use of the engineer's term ,"safety

factor." If a bridge is built to withstand a certain load without the beams permanently deforming, cracking, or breaking, it may be designed for the materials used to actually stand up under three times the load. This "safety factor" is to allow for uncertain excesses of load, or unknown extra loads, or weaknesses in the material that might have unexpected flaws, etc. If now the expected load comes on to the new bridge and a crack appears in a beam, this is a failure of the design. There was no safety factor at all; even though the bridge did not actually collapse because the crack went only one-third of the way through the beam. The O-rings of the Solid Rocket Boosters were not designed to erode. Erosion was a clue that something was wrong. Erosion was not something from which safety can be inferred.

There was no way, without full understanding, that one could have confidence that conditions the next time might not produce erosion three times more severe than the time before. Nevertheless, officials fooled themselves into thinking they had such understanding and confidence, in spite of the peculiar variations from case to case. A mathematical model was made to calculate erosion. This was a model based not on physical understanding but on empirical curve fitting. To be more detailed, it was supposed a stream of hot gas impinged on the O-ring material, and the heat was determined at the point of stagnation (so far, with reasonable physical, thermodynamic laws). But to determine how much rubber eroded it was assumed this depended only on this heat by a formula suggested by data on a similar material. A logarithmic plot suggested a straight line, so it was supposed that the erosion varied as the .58 power of the heat, the .58 being determined by a nearest fit. At any rate, adjusting some other numbers, it was determined that the model agreed with the erosion (to depth of one-third the radius of the ring). There is nothing much so wrong with this as believing the answer! Uncertainties appear everywhere. How strong the gas stream might be was unpredictable, it depended on holes formed in the putty. Blow-by showed that the ring might fail even though not, or only partially eroded through. The empirical formula was known to be uncertain, for it did not go directly through the very data points by which it was determined. There were a cloud of points some twice above, and some twice below the fitted curve, so erosions twice predicted were reasonable from that cause alone. Similar uncertainties surrounded the other constants in the formula, etc., etc. When using a mathematical model careful attention must be given to uncertainties in the model.

Liquid Fuel Engine (SSME)

During the flight of 51-L the three Space Shuttle Main Engines all worked perfectly, even, at the last moment, beginning to shut down the engines as the fuel supply began to fail. The question arises, however, as to whether, had it failed, and we were to investigate it in as much detail as we did the Solid Rocket Booster, we would find a similar lack of attention to faults and a deteriorating reliability. In other words, were the organization weaknesses that contributed to the accident confined to the Solid Rocket Booster sector or were they a more general characteristic of NASA? To that end the Space Shuttle Main Engines and the avionics were both investigated. No similar study of the Orbiter, or the External Tank were made.

The engine is a much more complicated structure than the Solid Rocket Booster, and a great deal more detailed engineering goes into it. Generally, the engineering seems to be of high quality and apparently considerable attention is paid to deficiencies and faults found in operation.

The usual way that such engines are designed (for military or civilian aircraft) may be called the component system, or bottom-up design. First it is necessary to thoroughly understand the properties and limitations of the materials to be used (for turbine blades, for example), and tests are begun in experimental rigs to determine those. With this knowledge larger component parts (such as bearings) are designed and tested individually.

As deficiencies and design errors are noted they are corrected and verified with further testing. Since one tests only parts at a time these tests and modifications are not overly expensive. Finally one works up to the final design of the entire engine, to the necessary specifications. There is a good chance, by this time that the engine will generally succeed, or that any failures are easily isolated and analyzed because the failure modes, limitations of materials, etc., are so well understood. There is a very good chance that the modifications to the engine to get around the final difficulties are not very hard to make, for most of the serious problems have already been discovered and dealt with in the earlier, less expensive, stages of the process.

The Space Shuttle Main Engine was handled in a different manner, top down, we might say. The engine was designed and put together all at once with relatively little detailed preliminary study of the material and components. Then when troubles are found in the bearings, turbine blades, coolant pipes, etc., it is more expensive and difficult to discover the causes and make changes. For example, cracks have been found in the turbine blades of the high pressure oxygen turbopump. Are they caused by flaws in the material, the effect of the oxygen atmosphere on the properties of the material, the thermal stresses of startup or shutdown, the vibration and stresses of steady running, or mainly at some resonance at certain speeds, etc.? How long can we run from crack initiation to crack failure, and how does this depend on power level? Using the completed engine as a test bed to resolve such questions is extremely expensive. One does not wish to lose an entire engine in order to find out where and how failure occurs. Yet, an accurate knowledge of this information is essential to acquire a confidence in the engine reliability in use. Without detailed understanding, confidence can not be attained.

A further disadvantage of the top-down method is that, if an understanding of a fault is obtained, a simple fix, such as a new shape for the turbine housing, may be impossible to implement without a redesign of the entire engine.

The Space Shuttle Main Engine is a very remarkable machine. It has a greater ratio of thrust to weight than any previous engine. It is built at the edge of, or outside of, previous engineering experience. Therefore, as expected, many different kinds of flaws and difficulties have turned up. Because, unfortunately, it was built in the top-down manner, they are difficult to find and fix. The design aim of a lifetime of 55 missions equivalent firings (27,000 seconds of operation, either in a mission of 500 seconds, or on a test stand) has not been obtained. The engine now requires very frequent maintenance and replacement of important parts, such as turbopumps, bearings, sheet metal housings, etc. The high-pressure fuel turbopump had to be replaced every three or four mission equivalents (although that may have been fixed, now) and the high pressure oxygen turbopump every five or six. This is at most ten percent of the original specification. But our main concern here is the determination of reliability.

In a total of about 250,000 seconds of operation, the engines have failed seriously perhaps 16 times. Engineering pays close attention to these failings and tries to remedy them as quickly as possible. This it does by test studies on special rigs experimentally designed for the flaws in question, by careful inspection of the engine for suggestive clues (like cracks), and by considerable study and analysis. In this way, in spite of the difficulties of top-down design, through hard work, many of the problems have apparently been solved.

A list of some of the problems follows. Those followed by an asterisk (*) are probably solved:

1. Turbine blade cracks in high pressure fuel turbopumps (HPFTP). (May have been solved.)
2. Turbine blade cracks in high pressure oxygen turbopumps (HPOTP).
3. Augmented Spark Igniter (ASI) line rupture.*
4. Purge check valve failure.*
5. ASI chamber erosion.*
6. HPFTP turbine sheet metal cracking.
7. HPFTP coolant liner failure.*
8. Main combustion chamber outlet elbow failure.*
9. Main combustion chamber inlet elbow weld offset.*
10. HPOTP subsynchronous whirl.*
11. Flight acceleration safety cutoff system (partial failure in a redundant system).*
12. Bearing spalling (partially solved).
13. A vibration at 4,000 Hertz making some engines inoperable, etc.

Many of these solved problems are the early difficulties of a new design, for 13 of them occurred in the first 125,000 seconds and only three in the second 125,000 seconds. Naturally, one can never be sure that all the bugs are out, and, for some, the fix may not have addressed the true cause. Thus, it is not unreasonable to guess there may be at least one surprise in the next 250,000 seconds, a probability of 1/500 per engine per mission. On a mission there are three engines, but some accidents would possibly be contained, and only affect one engine. The system can abort with only two engines. Therefore let us say that the unknown suprises do not, even of themselves, permit us to guess that the probability of mission failure do to the Space Shuttle Main Engine is less than 1/500. To this we must add the chance of failure from known, but as yet unsolved, problems (those without the asterisk in the list above). These we discuss below. (Engineers at Rocketdyne, the manufacturer, estimate the total probability as 1/10,000. Engineers at marshal estimate it as 1/300, while NASA management, to whom these engineers report, claims it is 1/100,000. An independent engineer consulting for NASA thought 1 or 2 per 100 a reasonable estimate.)

The history of the certification principles for these engines is confusing and difficult to explain. Initially the rule seems to have been that two sample engines must each have had twice the time operating without failure as the operating time of the engine to be certified (rule of 2x). At least that is the FAA practice, and NASA seems to have adopted it, originally expecting the certified time to be 10 missions (hence 20 missions for each sample). Obviously the best engines to use for comparison would be those of greatest total (flight plus test) operating time -- the so-called "fleet leaders." But what if a third sample and several others fail in a short time? Surely we will not be safe because two were unusual in lasting longer. The short time might be more representative of the real possibilities, and in the spirit of the safety factor of 2, we should only operate at half the time of the short-lived samples.

The slow shift toward decreasing safety factor can be seen in many examples. We take that of the HPFTP turbine blades. First of all the idea of testing an entire engine was abandoned. Each engine number has had many important parts (like the turbopumps themselves) replaced at frequent intervals, so that the rule must be shifted from engines to components. We accept an HPFTP for a certification time if two samples have each run successfully for twice that time (and of course, as a practical matter, no longer insisting that this time be as large as 10 missions). But what is "successfully?" The FAA calls a turbine blade crack a failure, in order, in practice, to really provide a safety factor greater than 2. There is some time that an engine can run between the time a crack originally starts until the time it has

grown large enough to fracture. (The FAA is contemplating new rules that take this extra safety time into account, but only if it is very carefully analyzed through known models within a known range of experience and with materials thoroughly tested. None of these conditions apply to the Space Shuttle Main Engine.

Cracks were found in many second stage HPFTP turbine blades. In one case three were found after 1,900 seconds, while in another they were not found after 4,200 seconds, although usually these longer runs showed cracks. To follow this story further we shall have to realize that the stress depends a great deal on the power level. The Challenger flight was to be at, and previous flights had been at, a power level called 104% of rated power level during most of the time the engines were operating. Judging from some material data it is supposed that at the level 104% of rated power level, the time to crack is about twice that at 109% or full power level (FPL). Future flights were to be at this level because of heavier payloads, and many tests were made at this level. Therefore dividing time at 104% by 2, we obtain units called equivalent full power level (EFPL). (Obviously, some uncertainty is introduced by that, but it has not been studied.) The earliest cracks mentioned above occurred at 1,375 EFPL.

Now the certification rule becomes "limit all second stage blades to a maximum of 1,375 seconds EFPL." If one objects that the safety factor of 2 is lost it is pointed out that the one turbine ran for 3,800 seconds EFPL without cracks, and half of this is 1,900 so we are being more conservative. We have fooled ourselves in three ways. First we have only one sample, and it is not the fleet leader, for the other two samples of 3,800 or more seconds had 17 cracked blades between them. (There are 59 blades in the engine.) Next we have abandoned the 2x rule and substituted equal time. And finally, 1,375 is where we did see a crack. We can say that no crack had been found below 1,375, but the last time we looked and saw no cracks was 1,100 seconds EFPL. We do not know when the crack formed between these times, for example cracks may have formed at 1,150 seconds EFPL. (Approximately 2/3 of the blade sets tested in excess of 1,375 seconds EFPL had cracks. Some recent experiments have, indeed, shown cracks as early as 1,150 seconds.) It was important to keep the number high, for the Challenger was to fly an engine very close to the limit by the time the flight was over.

Finally it is claimed that the criteria are not abandoned, and the system is safe, by giving up the FAA convention that there should be no cracks, and considering only a completely fractured blade a failure. With this definition no engine has yet failed. The idea is that since there is sufficient time for a crack to grow to a fracture we can insure that all is safe by inspecting all blades for cracks. If they are found, replace them, and if none are found we have enough time for a safe mission. This makes the crack problem not a flight safety problem, but merely a maintenance problem.

This may in fact be true. But how well do we know that cracks always grow slowly enough that no fracture can occur in a mission? Three engines have run for long times with a few cracked blades (about 3,000 seconds EFPL) with no blades broken off.

But a fix for this cracking may have been found. By changing the blade shape, shot-peening the surface, and covering with insulation to exclude thermal shock, the blades have not cracked so far.

A very similar story appears in the history of certification of the HPOTP, but we shall not give the details here.

It is evident, in summary, that the Flight Readiness Reviews and certification rules show a deterioration for some of the problems of the Space Shuttle Main Engine that is closely analogous to the deterioration seen in the rules for the Solid Rocket Booster.

Avionics

By "avionics" is meant the computer system on the Orbiter as well as its input sensors and output actuators. At first we will restrict ourselves to the computers proper and not be concerned with the reliability of the input information from the sensors of temperature, pressure, etc., nor with whether the computer output is faithfully followed by the actuators of rocket firings, mechanical controls, displays to astronauts, etc.

The computer system is very elaborate, having over 250,000 lines of code. It is responsible, among many other things, for the automatic control of the entire ascent to orbit, and for the descent until well into the atmosphere (below Mach 1) once one button is pushed deciding the landing site desired. It would be possible to make the entire landing automatically (except that the landing gear lowering signal is expressly left out of computer control, and must be provided by the pilot, ostensibly for safety reasons) but such an entirely automatic landing is probably not as safe as a pilot controlled landing. During orbital flight it is used in the control of payloads, in displaying information to the astronauts, and the exchange of information to the ground. It is evident that the safety of flight requires guaranteed accuracy of this elaborate system of computer hardware and software.

In brief, the hardware reliability is ensured by having four essentially independent identical computer systems. Where possible each sensor also has multiple copies, usually four, and each copy feeds all four of the computer lines. If the inputs from the sensors disagree, depending on circumstances, certain averages, or a majority selection is used as the effective input. The algorithm used by each of the four computers is exactly the same, so their inputs (since each sees all copies of the sensors) are the same. Therefore at each step the results in each computer should be identical. From time to time they are compared, but because they might operate at slightly different speeds a system of stopping and waiting at specific times is instituted before each comparison is made. If one of the computers disagrees, or is too late in having its answer ready, the three which do agree are assumed to be correct and the errant computer is taken completely out of the system. If, now, another computer fails, as judged by the agreement of the other two, it is taken out of the system, and the rest of the flight canceled, and descent to the landing site is instituted, controlled by the two remaining computers. It is seen that this is a redundant system since the failure of only one computer does not affect the mission. Finally, as an extra feature of safety, there is a fifth independent computer, whose memory is loaded with only the programs of ascent and descent, and which is capable of controlling the descent if there is a failure of more than two of the computers of the main line four.

There is not enough room in the memory of the main line computers for all the programs of ascent, descent, and payload programs in flight, so the memory is loaded about four time from tapes, by the astronauts.

Because of the enormous effort required to replace the software for such an elaborate system, and for checking a new system out, no change has been made to the hardware since the system began about fifteen years ago. The actual hardware is obsolete; for example, the memories are of the old ferrite core type. It is becoming more difficult to find manufacturers to supply such old-fashioned computers reliably and of high quality. Modern computers are very much more reliable, can run much faster, simplifying circuits, and allowing more to be done, and would not require so much loading of memory, for the memories are much larger.

The software is checked very carefully in a bottom-up fashion. First, each new line of code is checked, then sections of code or modules with special functions are verified. The scope is increased step by step until the new changes are incorporated into a complete system and checked. This complete output is considered the final product, newly released. But completely independently there is an independent verification group, that takes an adversary attitude to the software development group, and tests and verifies the software as if it were a customer of the delivered product. There is additional verification in using the new programs in simulators, etc. A discovery of an error during verification testing is considered very serious, and its origin studied very carefully to avoid such mistakes in the future. Such unexpected errors have been found only about six times in all the programming and program changing (for new or altered payloads) that has been done. The principle that is followed is that all the verification is not an aspect of program safety, it is merely a test of that safety, in a non-catastrophic verification. Flight safety is to be judged solely on how well the programs do in the verification tests. A failure here generates considerable concern.

To summarize then, the computer software checking system and attitude is of the highest quality. There appears to be no process of gradually fooling oneself while degrading standards so characteristic of the Solid Rocket Booster or Space Shuttle Main Engine safety systems. To be sure, there have been recent suggestions by management to curtail such elaborate and expensive tests as being unnecessary at this late date in Shuttle history. This must be resisted for it does not appreciate the mutual subtle influences, and sources of error generated by even small changes of one part of a program on another. There are perpetual requests for changes as new payloads and new demands and modifications are suggested by the users. Changes are expensive because they require extensive testing. The proper way to save money is to curtail the number of requested changes, not the quality of testing for each.

One might add that the elaborate system could be very much improved by more modern hardware and programming techniques. Any outside competition would have all the advantages of starting over, and whether that is a good idea for NASA now should be carefully considered.

Finally, returning to the sensors and actuators of the avionics system, we find that the attitude to system failure and reliability is not nearly as good as for the computer system. For example, a difficulty was found with certain temperature sensors sometimes failing. Yet 18 months later the same sensors were still being used, still sometimes failing, until a launch had to be scrubbed because two of them failed at the same time. Even on a succeeding flight this unreliable sensor was used again. Again reaction control systems, the rocket jets used for reorienting and control in flight still are somewhat unreliable. There is considerable redundancy, but a long history of failures, none of which has yet been extensive enough to seriously affect flight. The action of the jets is checked by sensors, and, if they fail to fire the computers choose another jet to fire. But they are not designed to fail, and the problem should be solved.

Conclusions

If a reasonable launch schedule is to be maintained, engineering often cannot be done fast enough to keep up with the expectations of originally conservative certification criteria designed to guarantee a very safe vehicle. In these situations, subtly, and often with apparently logical arguments, the criteria are altered so that flights may still be certified in time. They therefore fly in a relatively unsafe condition, with a chance of failure of the order of a percent (it is difficult to be more accurate).

Official management, on the other hand, claims to believe the probability of failure is a thousand times less. One reason for this may be an attempt to assure the government of NASA perfection and success in order to ensure the supply of funds. The other may be that they sincerely believed it to be true, demonstrating an almost incredible lack of communication between themselves and their working engineers.

In any event this has had very unfortunate consequences, the most serious of which is to encourage ordinary citizens to fly in such a dangerous machine, as if it had attained the safety of an ordinary airliner. The astronauts, like test pilots, should know their risks, and we honor them for their courage. Who can doubt that McAuliffe was equally a person of great courage, who was closer to an awareness of the true risk than NASA management would have us believe?

Let us make recommendations to ensure that NASA officials deal in a world of reality in understanding technological weaknesses and imperfections well enough to be actively trying to eliminate them. They must live in reality in comparing the costs and utility of the Shuttle to other methods of entering space. And they must be realistic in making contracts, in estimating costs, and the difficulty of the projects. Only realistic flight schedules should be proposed, schedules that have a reasonable chance of being met. If in this way the government would not support them, then so be it. NASA owes it to the citizens from whom it asks support to be frank, honest, and informative, so that these citizens can make the wisest decisions for the use of their limited resources.

For a successful technology, reality must take precedence over public relations, for nature cannot be fooled.