DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES

CALIFORNIA INSTITUTE OF TECHNOLOGY

PASADENA, CALIFORNIA 91125

KEEPING SCORE: AN ACTUARIAL APPROACH TO ZERO-INFINITY DILEMMAS

Talbot Page

SOCIAL SCIENCE WORKING PAPER 248

January 1979

KEEPING SCORE: AN ACTUARIAL APPROACH TO ZERO-INFINITY DILEMMAS

Introduction

A "zero-infinity dilemma", where there is presumed to be a very low ("nearly zero") probability of a very high ("nearly infinite") cost, is the least tractible yet most important type of risk problem associated with evaluating prospective energy alternatives. The term has often been used to describe the nature of the risk associated with nuclear power and many of the best known examples have been drawn from this field, but this type of risk problem is associated with the other main energy alternatives as well. In the incomplete combustion of fossil fuels, for just one example, large scale release of aromatic hydrocarbons poses another zero-infinity dilemma, with the risk of cancer and mutation.

The obvious difficulty in dealing with zero-infinity dilemmas is that there is little or no direct historical record to develop probability estimates based upon frequency. By way of illustration, in its summary the Rasmussen report predicts that there is a one in five billion chance, per reactor year, of a catastrophic core meltdown. But the direct historical record of so many hundred reactor years with no meltdown is not greatly reassuring. This negative information, however welcome, is compatible with a one in a thousand chance of catastrophic meltdown as well as the one-in-five-billion Rasmussen estimate. The actuarial record, taken in simple terms of meltdowns per reactor year, is far too short to discriminate between the two. Nor do we want to wait for a long record of history of reactor years in order to develop an estimate of the probability of meltdown. On the contrary, we need the estimate of probability before a large scale program is undertaken, in order to decide upon the safety features and the scale of the program.

Decision Trees and Accountability

In spite of this somewhat pessimistic introduction, I believe that an actuarial approach, broadly conceived, has something constructive to say about the more accurate quantification of risk even in the toughest case of the zero-infinity dilemma. We may begin with a comment by William Stephenson, director of British Intelligence in World War II: "I can speak only of industrial 'accidents'. If you have access to insurance company files, you will see detailed studies of the weak point in any manufacturing process or mining procedure. Insurance companies stand to lose fortunes from an accident, and so they employ experts to figure out every possible way that things can go wrong. Their reports are guidebooks for saboteurs."[1]

Two principal observations follow from this comment. First, the comment suggests that decades ago insurance companies were using, at least informally, an approach of error analysis which later flowered in the enormous complexity of the Rasmussen report and similar treatments of decision trees. From this observation it may seem curious that as the procedures of risk assessment have become greatly more sophisticated, the public's level of confidence in the risk assessments, especially those related to nuclear power, appears to have declined. This phenomenon has led some practitioners of risk assessment to suggest that perhaps the public is irrational, not understanding the techniques of risk assessment, in particular in dealing with low probabilities. But, if this perceived phenomenon is a real one, and public skepticism has actually increased, the second observation suggests that this skepticism may in fact be rationally based once the nature of the incentive structure is taken into account.

In dealing with incentives relating to insurance, economists

have focused upon the behavior of the insured party and the problem of "moral hazard". As a second observation Stephenson reminds us that we should also consider the incentive structure upon the risk assessors themselves. In the normal practice of insurance (which does not deal with societal zero-infinity dilemmas) an insurance company can find out relatively quickly how accurately it has assessed the probabilities of various accidents; moreover, it stands to lose greatly if its assessments are inaccurate. The situation is different when there are zero-infinity dilemmas.

Most zero-infinity dilemmas appear to share a number of other characteristics besides the presumed low probability and the potentially catastrophic outcome.[2] In our discussion here the most important other characteristic is latency. For the greenhouse effect associated with fossil energy sources, it takes many years for CO_2 to build up and the projected climate changes to occur and decades to know whether or not there would be catastrophic effects following from our present decisions about fossil fuel. There are similarly long latency periods between the initiation of action and the manifestation of result from chemical carcinogens or radiation. Even such a dramatic event as a core meltdown is associated with a long latency in terms of the expected waiting time for its first occurrence. Latency means that there are few, if any, direct tests of the risk assessor's accuracy. Lack of a direct test means a lack of accountability, unless of course indirect tests can be devised.

Latency, by no means only to be associated with the zero-infinity dilemmas, explains a common pattern of new industries. In the first stage, it is to the advantage of the industry development to understate the risks and other costs associated with it. This incentive applies to all types of risks

and costs but with special force to risks of the zero-infinity character because of the greater latency before actual knowledge of costs and risks. Certain alternatives, for example microwave transmission from satellite collection of solar energy, appear to be in this first stage. I am not suggesting here that the current assessments of risk and cost are in fact underassessments, but that at early stages of a new industry there are incentives, stemming from latencies of adverse effects, for such underassessment.

After a few decades, if several of the more optimistic predictions have proved false, a generalized skepticism is likely to set in, articulated in large part as observations about incentives upon the risk and cost assessors. It appears that the nuclear industry has moved into this second stage. One often hears such statements as "I have little faith in so-and-so's estimate, he consults for the industry", or "Of course he assesses the risks very low, he is a nuclear engineer and his career is at stake". While such ad hominem interpretations are no doubt galling to the risk assessors involved, protestations of objectivity have little impact at this stage. In the second stage the principal concern is establishing, or reestablishing, the credibility of the risk and cost assessors. Again I wish to emphasize that there is no necessary connection between an incentive and an outcome. But the public is right to pay attention to incentives.

Careful attention to institutional design and the role of incentives in it is beneficial for the following reasons: increasing the accountability of the risk assessors, by developing more explicit direct and indirect tests of their predictions, is the most constructive way of increasing public confidence in the entire assessment process; at the same time better assessments of risk may be obtained. Separating safety regulation from development in the AEC is an example of attention to incentives.

Keeping Score

In sporting events score is kept at two levels. At one level keeping score is a measure of the individual's or team's effort or success of effort, for example the number of goals scored in the first ten minutes of a basketball game. On the second level score is kept of the wins and losses of an individual in competition with others. Even where there is no formalized competition, score as a measure of effort is kept, for example in solo golf or in first ascents in mountain climbing, and this measure of effort is a requisite for scoring wins and losses in a formal, competitive setting. Still it is the scoring of the wins and losses that counts in most sports.

In the field of risk assessment, we are familiar with keeping score as a measure of the assessor's efforts, but less familiar with scoring the wins and losses of one assessor in direct competition with another. And indeed many may consider this an undesirable direction to go in, believing that the "game" of risk assessment is already too adversarial, especially in the area of nuclear energy. However, it is worth noting that the actuarial approach, in the setting of insurance industries, is directly competitive. One company competes against another, and the market keeps score of wins and losses.

In the area of energy, risk assessment, especially for zero-infinity dilemmas, is mainly a government sponsored activity. But even if governments wished to establish competitive tests of performance, in some sense analogous to market tests, the task is made difficult by latency and the infrequency of observation of rare events. There are occasions, however, where latency is not a problem or where indirect tests could be constructed. It is interesting to speculate upon how a formalized competitive approach might go.

Consider the partly hypothetical example of the SNAP reactor, which was developed as a power plant for space vehicles. Before the satellite was launched the AEC predicted a very low probability of failure. An assessment of a 10^{-8} probability of failure appeared, although this was probably an informal and not official estimate. During the same period skeptics assessed the risk of failure to be much higher. While I doubt that a numerical estimate was produced that reflected the consensus view of the skeptics, or official estimate of a known organization of skeptics, such as the Natural Resources Defense Council, for the sake of illustration we may attribute an assessment of 10^{-4} probability of failure to NRDC.

We can imagine that a game is being played, with the immediate object of increasing one's credibility. We begin the game with an initial distribution of credibility among the players. On the basis of the relative sizes of the organizations and the numbers of experts in each, suppose we assign an initial credibility measure of .99 to the AEC and 0.01 to NRDC. The players gain or lose credibility depending on their risk assessments and the outcome of the experiment, in this case the launching of the satellite. We define two states:

Θ_{AEC} The AEC is accurate in its risk assessment

Θ_{NRDC} NRDC is accurate in its risk assessment;

the two risk assessments for the SNAP reactor failure:

$$P(F|\Theta_{AEC}) = 10^{-8}$$

$$P(F|\Theta_{NRDC}) = 10^{-4}.$$

When there are many experiments, or rounds, we can write

$P^t(F|\Theta_{AEC})$ AEC assessment of probability of failure in round t;

$P^t(F|\Theta_{NRDC})$ NRDC assessment of probability of failure in round t;

$P^t(S|\Theta_{AEC})$ AEC assessment of probability of success in round t;

$P^t(S|\Theta_{NRDC})$ NRDC assessment of probability of success in round t.

We start with the two initial measures of credibility

$$C(0, AEC) = 0.99;$$

$$C(0, NRDC) = 0.01;$$

these initial values correspond to poll positions in a horse race.

We can define the credibility measures in round t by a formula corresponding to Bayes' Theorem:

If round t turns out to be a failure

$$(2) \qquad C(t,AEC) = \frac{P^t(F|\Theta_{AEC})C(t-1, AEC)}{P^t(F|\Theta_{AEC})C(t-1, AEC) + P^t(F|\Theta_{NRDC})C(t-1, NRDC)}.$$

And if round t turns out to be a success

$$C(t,AEC) = \frac{P^t(S|\Theta_{AEC})C(t-1, AEC)}{P^t(S|\Theta_{AEC})C(t-1, AEC) + P^t(S|\Theta_{NRDC})C(t-1, NRDC)},$$

and

$$C(t, NRDC) = 1 - C(t, AEC)$$

In the second round we would start with credibility measures (C(1,AEC),C(1,NRDC)), and the contested risk assessments for some other event, which could be a second try for the SNAP reactor or a quite different event whose rarity is at issue between the AEC and NRDC, but which can be observed and scored. Then we would compute the second round credibilities from (1) with t = 2. The initial credibility measure (C(0,AEC),C(0,NRDC) is the same as the prior distribution in decision analysis; and the first round credibility measure (C(1,AEC),C(1,NRDC)) is the same as the posterior distribution. It would not be worth recasting the notation from that in standard decision theory, except for the slight extension from the usual interpretation -- it is assumed that Θ_{AEC} refers to an entire technique and capacity for risk analysis, and thus the credibility measure describing the AEC assessment of a SNAP failure also describes AEC assessments of other risks.[3]

In the case of the SNAP reactor, the experiment was undertaken and unfortunately the reactor failed upon its first try. Direct computation from Bayes' Theorem shows that the credibility measure for the AEC falls from 0.99 to 0.01 and the measure for NRDC rises correspondingly. Such an enormous fall in credibility is startling and it raises a number of interesting issues.

First of all it illustrates that the initial seeding of players, which is done on the basis of outside judgment, can be quickly overcome by an actual track record. On the contrary we might ask should a credibility measure be defined to be capable of such volatility? This is really a question of the applicability of standard decision theory to low probability events, as it has nothing to do with our assumption of extension, which says that the credibility measure applies to the entire risk assessment capacity, not just one prediction, and allows us to go from one round to the next. This question has to do with the relative weights to be given to the prior and the new evidence, which is a basic question in decision analysis.

Without deeper discussion of this last question, we may note that the volatility is asymmetric. If the SNAP reactor had succeeded NRDC's credibility measure would have fallen, but its percentage change would have been much smaller. At least it can be said that the asymmetry is in the right direction. When both parties agree that a failure is rare, there is more information in a failure than in a success (as to the accuracy of the risk assessments).

If we look at this credibility "game" as a sequential process lasting hundreds or thousands of rounds, where not all the probability assessments have to be on rare events, we may ask questions of convergence. Suppose that the AEC were indeed more accurate in its risk assessment than NRDC. How long would it take to show up in its credibility measure? For a long sequence of rounds, how much does the initial posting or seeding of players matter? Given that the object of the game is to maximize one's own credibility over a sequence of rounds, is a player better off publishing his own "sincere"

estimates of risk or other numbers depending on what he thinks the other player is going to do? In other words, are there dominant strategies and if so, do they produce unbiased estimates of risk?

If credibility games could be developed with sincere dominant strategies, they could provide powerful incentives toward more accurate risk assessment. The incentive would apply to both players, so that one would expect a convergence of risk assessments, a welcome phenomenon. An important purpose of establishing a formal game of the sort sketched above is to provide for near term accountability for both players, for assessments and statements about their assessments. At the present time it is unclear to the outsider how seriously to take the failure of the SNAP reactor, for example. It is possible that the AEC has made hundreds of thousands of predictions and NRDC has seized upon the unlucky one. On the other hand it is possible that the AEC got away with overly optimistic predictions by not making them formal and then keeping quiet about the failures. Without keeping score on the wins and losses there is no way to tell.

Even though at present there is no accountability of the risk assessors along the lines of the formal game outlined above, to some extent this game informally exists already (and in a marred form). Agencies lose credibility when they assign very low probabilities to events which then happen, especially when critics have assigned larger probabilities. Agencies gain credibility in the reverse situation. Thus it would be useful to analyze the winning strategies in such games to see what incentives are on the players, even in the informal situation.

As mentioned earlier, even without establishing a formal competition and scoring the wins and losses of the players in comparison with each

other, it is possible to keep score on the individual level, of the solo player. When we are dealing with final events, such as the success or failure of a reactor, with very low assessed probabilities of failure, just one or two failures casts doubt upon the assessment process, unless there are hundreds of millions of low probability assessments being made.

A comparison can be made with the problem of assessing the probability of a perfect bridge hand (where one player gets all clubs, another all hearts, etc.). Textbook exercises, based upon combinatorial principles, yield an extremely low estimate of probability. The estimate is so low that a single occurrence of a perfect hand casts doubt on the probability estimate. And as perfect hands are reported every few years, the assessment technique itself has been reconsidered. Upon review it appears that some perfect hands are the result of gags (which would correspond to Stephenson's sabotage). Moreover, it appears that when sabotage is prevented, as in tournaments, the probability of a perfect hand is still too high to be explained by a combinatorial calculation. It appears much more likely that perfect hands result from a "common mode failure" -- that is, the probabilities of one intermediate event are conditional upon another intermediate event. New decks come in perfect hands and shuffling is not entirely random, so that dealing is not made up of independent events. Thus even the single occurrence of a presumed very rare event can lead to the reassessment of the risk assessment technique and the adoption of more realistic techniques.

The evaluation of risk assessment is made more difficult by the fact that many of the gambles being assessed are not repeatable experiments. They are assessments of gambles about the state of nature, which may be one way or another. Each gamble is assessed at its own probability, typically a very low one for the kinds of problems we are considering, and then scored

0 or 1, depending on whether the benign or adverse hypothesis about the state of nature turns out to be true. It is not possible to tell whether or not a single assessment of the likelihood of a particular hypothesis is realistic or not. However, with the assumption of extension (that an agency's assessment technique is consistently accurate or inaccurate) it is still possible to score an overall effort.

Suppose that an agency were totally accurate in its risk assessments and its estimate of the risk of failure (or adverse hypothesis) of the ith gamble is p_i. Then the sum of N such Bernoulli random variables has expectation

$$\sum_{i=1}^{N} p_i$$

and variance $\sum p_i(1-p_i)$.

This binominal-like random variable (the p_i are different) allows us to score the aggregate effort of the risk assessor and ask how likely it is that the assessor has been accurate.

Because of latency, there is limited opportunity to develop this type of scorecard for a solo player, if we insist upon evaluating probability estimates of final events, such as core meltdowns. However, this technique applies to the scoring of intermediate events and partial chains upon fault trees. In fact it may apply more suitably to the evaluation of probability assessments of intermediate events than final events. Not only are there more intermediate events to evaluate, often with less of a latency problem, but also evaluation of intermediate events lends itself to testing assumptions of independence. Common mode failures take place in the intermediate chains of events more directly than across final events. As the binominal-like distribution assumes independence of the underlying Bernoulli random variables,

this assumption is tested simultaneously with the assumption of accuracy of the individual assessment of the p_i.[4]

The binominal-like distribution provides a way of testing the aggregative performance of the risk assessor. Depending on the types of assessments being evaluated it can also test as well the assessor's allowance for, or treatment of, the possibilities of human error (e.g. the initiation of the Brown's Ferry accident) or his own tunnel vision (leaving out paths or focusing on the wrong ones).

Conclusion

Interpreted narrowly as the method of calculating probabilities of final events by their historical frequency, the actuarial approach has limited application for risk analysis in the energy area, especially where there are low probabilities of catastrophes, and often latencies. Interpreted broadly as the approach to risk estimation found in insurance markets, the actuarial approach has much to offer. The actuarial approach, in the context of market competition, uses the historical record to reward the companies with accurate risk assessors and starve the others. The actuarial approach directs us to increase the accountability of the risk assessors in government sponsored assessments and keep score of the success of effort of the individual risk assessors. The approach suggests an analysis of the incentive structure on the assessors themselves, and of the strategies of the (informal) games the assessors are placed into. The goals of the actuarial approach, broadly conceived, are (1) to achieve better point estimates of probabilities, (2) to achieve a better understanding of the credibility or subjective confidence interval to be associated with the point estimates, (3) to understand better the incentives for under- or overassessment, and (4) to increase the level of confidence of the public in the entire assessment process, through explicit evaluation of performance.

NOTES

1. William Stevenson, A Man Called Intrepid, Ballantine Books: New York 1976, p. 64.

2. For a discussion of nine characteristics of this type of risk see Talbot Page, "A Generic View of Toxic Chemicals and Similar Risks," Ecology Law Quarterly, symposium issue on toxic chemicals, 1978.

3. For an application of decision theory without the extension assumption see R.A. Howard, J.E. Matheson, and D.W. North, "The Decision to Seed Hurricanes," Science, Vol. 176, No. 4040 (16 June 1972), pp. 1191-1202. In this discussion the prior and posterior distributions can be interpreted as original and revised credibility measures on the hypotheses put forward by differing scientists. Without the assumption of an extended and roughly constant capacity of risk assessment, there is just one round.

4. The assumption of extension -- that the accuracy ability of an individual risk assessor is roughly constant across experiments -- appears to be the underlying principle in a series of insightful experiments on the evaluation of risk assessors by Alpert and Raiffa and by Tversky and Kahneman. These investigators used a list of "almanac" questions such as: "How many foreign cars were imported into the United States in 1968?" For each question the test subjects were asked to make an estimate so high that they would believe there was only a one percent probability that the true answer would exceed their estimate, and similarly for a low estimate. Thus the subjects constructed rare events -- being outside their confidence intervals. Although there is no way of telling if a single estimate is accurate, under the assumption that the assessors were consistently accurate one would expect the factual answers to be outside the constructed intervals about two percent of the time, over a number of trials. The experiments suggested that the subjects were not consistently accurate assessors, because the factual answers were outside the constructed intervals 40 to 50 percent of the time. These experiments are described in Slovic, Kunreuther, and White, "Decision Processes, Rationality, and Adjustment to Natural Hazards," in Natural Hazards: Local, National, and Global (G. White ed.) 1974.