1

A Theory of Verification

Доверяй, но проверяй!

— Vladimir Ilyich Lenin

Let us begin with an observation in the spirit of the above: Verifying the fulfillment of commitments is a necessary evil, since people know from long experience that they cannot trust one another.

This is a book about that evil necessity, more specifically about its theoretical foundations and a rational basis for its practical implementation. We shall be dealing primarily with institutionalized, multilateral systems of verification, such as those arising from arms control agreements or treaties for the protection of the environment, and we shall attempt to treat problems arising from these areas in a quantitative, consistent way.

The Oxford English Dictionary defines verification as

> ... the action of establishing or testing the truth or correctness of a fact, theory, statement, etc., by means of special investigation or comparison of data.

Further refinements of such a clear definition are superfluous, although a number have been made. The United Nations Secretary-General in 1978, for example:

> ... the process of ascertaining that a commitment laid down on a particular agreement in the field of disarmament or arms limitation is being met.

Or, as the United States Arms Control and Disarmament Agency put it in its glossary to the 1979 SALT II Treaty:

> ... the process of determining to the extent necessary to safeguard national security adequately that the other side is complying with an agreement.

Kokoski (1990) gives in his introduction to an analysis of the Treaty on Conventional Forces in Europe a rather concise description of the objectives of verification:

> The most obvious [purpose] is to *detect* violations of an agreement, thereby to provide early warning to deny any advantage to a violator. The second purpose is to *deter* violations by the fact that verification increases the risk of detection. The third main purpose is to *build confidence*, not only among treaty partners but also within domestic political communities. Finally, verification aims to *clarify uncertainty*.

Krass (1985), Potter (1985) and the UN Secretary-General (1990), provide some recent qualitative treatments of the subject. For a discussion of verification in the specific context of arms control agreements see Calogero *et al.* (1990), Graybeard *et al.* (1991), Fischer (1991), Goldblat (1982) and many articles in The Bulletin of the Verification Technology Information Centre (VERTIC), London. The verification of environmental treaties (or lack of it) is discussed in Hajost and Shea (1990). Table 1.1, relegated to the end of this chapter, lists some existing arms control and environmental agreements and outlines their relevant verification provisions.

1.1 Cooperation versus Confrontation

Although negative reflections on human trustworthiness may well be justified, they should perhaps be tempered a little, since experience also teaches that most of us can trust one another most of the time. How, then, should a verification system be seen? As a cooperative effort in which each involved party freely volunteers proof of adherence to commitments, or as a confrontation between inspectee and inspector, with the latter distrusting the former as a matter of professional principle? The answer, confusingly, is yes! A workable verification regime calls for a high measure of cooperation to function at all, whereas credibility demands that the design and assessment of inspection procedures assume a *deliberate, planned attempt by the inspected party to behave illegally*. The latter requirement, being based upon a purely hypothetical assumption, doesn't necessarily contradict the former, though of course it may. How one can treat the confrontational side of the problem consistently is in fact the subject of this book.

But is a formal theory of verification really necessary? Again, yes! It is needed in order to demonstrate credibility, in order to design optimal inspection procedures, in order to apportion finite resources efficiently

and in order to achieve impartiality and objectivity in the assessment of effectiveness. Without a quantitative basis, such things are not possible.

Certainly there exists no rigorous, axiomatic theory of verification, desirable as that might be, since the problems encountered in reality are far too diverse. There are however prototypic problems which seem to arise over and over again in many different contexts and for which methods of analysis have been developed. These methods consist of a convenient marriage (as opposed to a marriage of convenience) of statistics and the mathematics of non-cooperative games. They are related to, but differ quite fundamentally from, the statistical techniques used in quality control problems. The difference arises out of the adversarial or confrontational aspect which, as we have said, is special to the problem and which must be taken into account if the theory is to be meaningful.

We shall be discussing what we hope to be illustrative and practical applications of verification theory in subsequent chapters, but, in order to get the flavor of things to come, let us first accompany an imaginary inspection team on a trip to an imaginary land (Canty and Avenhaus (1988)).

1.2 A Ruritanian Example

The peace-loving kingdom of Ruritania has signed the Convention on Chemical Weapons,[1] pledging neither to produce nor to acquire chemical weapons of mass destruction, and submitting its entire civilian chemical industry (consisting of a fertilizer factory and an insecticide plant) to routine inspection under the terms of the Convention.

Facility designs, production capacities and schedules, that were made known to the inspectorate and confirmed by initial visits, indicate that both plants might be capable of producing chemicals banned by the agreement. In the case of the fertilizer plant, the inspectorate has concluded that such misuse could be detected with certainty on subsequent annual inspections, whereas for the insecticide plant, only a 50-50 chance would exist for detection. The reader will now forgive us if we elect, for purely didactic reasons, to interpret these assessments literally, i.e. 100% and 50% detection probabilities for misuse of fertilizer and insecticide plants, respectively. The inspectorate is short of manpower, other countries having somewhat larger chemical industries, and would like to inspect the Ruritanian facilities on a random basis (Figure 1.1). Can it

[1] At this writing not yet in force. Implementation in Ruritania and elsewhere won't begin until 180 days after 65 signatory states have ratified the Convention; see Table 1.1.

do so without sacrificing detection capability? A theoretical question if ever there was one!

First of all, what is the inspectorate's detection capability if there is no randomization, in other words if both plants are inspected routinely? Here some realistic pessimism is called for. The potential culprit (the Ruritanian Minister for Defense and Agriculture) may be aware of the inspectorate's *a priori* detection capabilities, and therefore the inspectorate should assume that the fertilizer plant will *never* be misused when under regular routine inspection. The detection probability is thus 50%.

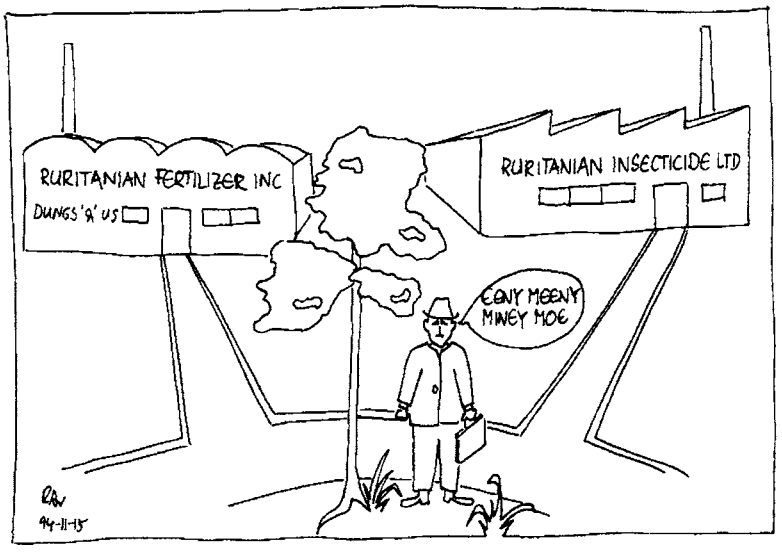

Fig. 1.1. An inspector in Ruritania.

Now consider the following randomization strategy: *Once a year, at a randomly chosen date, inspect the fertilizer plant with 50% probability; always inspect the insecticide plant.*

On average the inspectorate will save one half the resources normally needed for one inspection, and the operator of the fertilizer plant will be spared the nuisance of an inspection half the time. But what is the detection probability? Evidently it is *still* 50%. For if the insecticide plant is misused, it will be detected with 50% probability due to the nature of the inspection measures applied there, whereas misuse of the

fertilizer plant will be caught for sure if an inspection takes place, and this happens with a 50% probability.

All of which is no doubt painfully obvious, if not downright trivial. But let's pose some more theoretical questions. Suppose the detection probability at the fertilizer plant is only 90%. Can the inspectorate still maintain an overall detection probability of 50% under randomization with the same effort? Or suppose the sum of the inspection probabilities is 1.3 rather than 1.5. What is the worst case detection probability now? The answers to questions of this kind can be found in Chapter 6, where we generalize to arbitrarily many facilities, arbitrary *a priori* detection probabilities and allow for the possibility of false alarms. But for now let's stay in Ruritania and confirm our intuition with some simple mathematics.

Label the insecticide and fertilizer plants I and F, respectively, and denote the misuse detection probabilities $1-\beta_I$ and $1-\beta_F$. Here we've adopted the decision theorist's rather awkward convention of using the Greek letter β for the *non*-detection probability. For our specific example violations at the insecticide plant were harder to detect, so

$$0 \le 1-\beta_I < 1-\beta_F \le 1. \tag{1.1}$$

Similarly let the inspection probabilities be p_I and p_F, with

$$0 \le p_I, p_F \le 1, \quad p_I + p_F = k \le 2. \tag{1.2}$$

The parameter k gives the degree of randomization. If $k = 2$ there is no randomization since both plants are inspected with probability 1, while if $k = 0$ there are no inspections at all (which might be carrying things a bit too far). We can think of k as a rough measure of the total annual inspection effort for Ruritania. The condition (1.2) may be a little difficult to grasp. Perhaps it is easier to see as follows: The *expected* number of inspection visits is

$$\begin{aligned} 0 \cdot (1-p_I)\cdot(1-p_F) + 1 \cdot [(1-p_I)\cdot p_F + p_I \cdot (1-p_F)] + 2 \cdot p_I \cdot p_F \\ = p_I + p_F. \end{aligned} \tag{1.3}$$

Finally, we can define the hypothetical probabilities for facility misuse as q_I and q_F:

$$0 \le q_I, q_F \le 1, \quad q_I + q_F = 1. \tag{1.4}$$

It has been assumed, on the conservative side from the inspector's point

of view, that misuse will take place either in plant A or in plant B but not in both. Hence the sum of unity.[1]

The overall detection capability of the randomized inspection strategy is obviously

$$1-\beta(p_I,p_F;q_I,q_F)=p_I\cdot q_I\cdot(1-\beta_I)+p_F\cdot q_F\cdot(1-\beta_F) \tag{1.5}$$

or, using the equations (1.2) and (1.4) to eliminate q_F and p_F,

$$1-\beta(p_I;q_I)=p_I\cdot q_I\cdot(1-\beta_I)+(k-p_I)\cdot(1-q_I)\cdot(1-\beta_F). \tag{1.6}$$

So far so good, but there is as yet no way to determine p_I and q_I. Introducing a little game theoretical terminology, we shall call p_I^* the *inspector's optimal strategy* in ignorance of the violator's (illegal) intentions, and q_I^* the *violator's optimal strategy* similarly unaware of the inspector's intentions. These two optimal choices, once determined, can be put into equation (1.6) to give the *equilibrium detection probability*

$$1-\beta^* := 1-\beta(p_I^*;q_I^*). \tag{1.7}$$

What do we mean by *optimal strategy* and *equilibrium*? These ideas are quite simple and intuitive: If the inspectorate inspects plants I and F with optimal probabilities p_I^* and $p_F^*=k-p_I^*$, respectively, then *no matter what the violator does* he cannot force the detection probability below $1-\beta^*$. Conversely, if the violator misuses plants I or F with optimal probabilities q_I^* and $q_F^*=1-q_I^*$, respectively, then *no matter what the inspector does* he cannot achieve a detection probability higher than $1-\beta^*$. Therefore either protagonist would be ill-advised to deviate from such an optimal strategy. This situation can be expressed mathematically very concisely with the inequalities

$$1-\beta(p_I;q_I^*)\le 1-\beta(p_I^*;q_I^*)\le 1-\beta(p_I^*;q_I) \text{ for all } p_I,q_I,$$

which are called the *equilibrium* or *saddle point* criteria for a *two-person, zero-sum game* with payoff $1-\beta(p_I;q_I)$ to the inspector. In fact throughout this book we shall prefer to use the equivalent formulation

$$\beta(p_I^*;q_I)\le\beta(p_I^*;q_I^*)\le\beta(p_I;q_I^*) \text{ for all } p_I,q_I, \tag{1.8}$$

simply because it saves a bit of writing. A graphical representation of a

[1] Is a random violation strategy a reasonable hypothesis? Yes, in fact it is just as reasonable as a randomized inspection strategy. Try playing several repetitions of *Guess Which Hand's Got the Dollar* against any opponent over the age of 4 to appreciate the efficacy of randomization.

typical saddle point is shown in Chapter 4, Figure 4.2. Determining the optimal strategies is equivalent to 'solving' (1.8).

Whilst we might, for ethical reasons, hesitate to recommend the optimal misuse strategy to the Ruritanian Minister, we can strongly recommend the inspector's optimal strategy, as it *guarantees* him the equilibrium detection probability $1-\beta(p_I^*;q_I^*)$. In fact the adjectives *equilibrium, saddle point* and *guaranteed* will be used interchangeably when applied to detection probabilities. Because of this desirable property, we shall be concerned throughout this book with establishing equilibria.

But, the reader may protest, why zero-sum?! Surely there is more to life in general and treaty verification in particular than detection probabilities. The answer is given in Chapter 9 where it is shown that saddle point strategies are, in a much more general sense, optimal.

We can now write down the solution to the game, that is those values p_I^* and q_I^* that satisfy (1.2), (1.4) and (1.8). It turns out that the form of the solution is dependent on the effort parameter k. For

$$0 \le k \le 1 + \frac{1-\beta_I}{1-\beta_F} \tag{1.9}$$

the solution is

$$p_I^* = k \cdot \frac{1-\beta_F}{1-\beta_I+1-\beta_F}, \quad q_I^* = \frac{1-\beta_F}{1-\beta_I+1-\beta_F} \tag{1.10}$$

giving an equilibrium detection probability (1.6) of

$$1-\beta^* = k \cdot \frac{(1-\beta_I)\cdot(1-\beta_F)}{1-\beta_I+1-\beta_F} \tag{1.11}$$

which may be written less explicitly but more symmetrically as

$$\frac{k}{1-\beta^*} = \frac{1}{1-\beta_I} + \frac{1}{1-\beta_F}. \tag{1.12}$$

The solution for larger values of k, i.e.

$$1 + \frac{1-\beta_I}{1-\beta_F} \le k \le 2, \tag{1.13}$$

is even simpler. It is

$$p_I^* = 1, \quad q_I^* = 1 \tag{1.14}$$

with an equilibrium detection probability of

$$1-\beta^* = 1-\beta_I. \tag{1.15}$$

The above equations give a complete solution to the RRIP (Ruritanian Randomized Inspection Problem). They can be proved by showing that they satisfy the saddle point criteria (1.8), an easy exercise which the reader may care to try. But let's look at our solutions in more detail, and see if they make sense.

It can be seen that if k is not too large, satisfying (1.9), equality holds for both of the saddle point criteria (1.8). Each player behaves so as to make his opponent indifferent with respect to his strategy choice. Nevertheless the opponent's optimal strategy is uniquely defined! The optimal inspection probabilities p_I^* and p_F^* are in fact proportional to the total inspection effort k (see equation (1.10)) while their ratio is independent of k:

$$\frac{p_I^*}{p_F^*} = \frac{1-\beta_F}{1-\beta_I}.$$

The optimal violation probabilities are completely independent of k, but their ratio is the same, that is,

$$\frac{q_I^*}{q_F^*} = \frac{1-\beta_F}{1-\beta_I}.$$

This is reasonable: The larger the detection probability $1-\beta_F$ the smaller the probability of misuse of plant F, and therefore the smaller the probability for inspecting F as well.

If the total inspection probability or effort becomes large enough, solution (1.14–15) obtains. The violator will concentrate exclusively on plant I since there at least the probability of detection is smaller. As a consequence, the inspectorate will inspect I with certainty.

Note from (1.15) that the largest achievable detection probability for the inspectorate is $1-\beta^* = 1-\beta_I$ and that it can be attained when

$$k = k_0 = 1 + \frac{1-\beta_I}{1-\beta_F}.$$

Any inspection effort exceeding k_0 is wasted! The theory thus also recommends an upper value for the total annual inspection effort.

Returning to our original example, for which

$$1-\beta_I = 0.5, \quad 1-\beta_F = 1.0,$$

it follows from (1.14) and (1.15), with $k = 1.5$, that

$$1-\beta^* = 0.5, \quad p_I^* = 1.0, \quad p_F^* = 0.5, \quad q_I^* = 1.0, \quad q_F^* = 0.0$$

which was the intuitively obvious result. We then asked about the situation

$$1 - \beta_I = 0.5, \quad 1 - \beta_F = 0.9.$$

The value of k which still yields $1 - \beta^* = 1 - \beta_I$ is

$$1 + \frac{1 - \beta_I}{1 - \beta_F} = 1.556.$$

If the effort remains constant, i.e. $k = 1.5$, equations (1.10) and (1.11) are applicable and give

$$1 - \beta^* = 0.48, \quad p_I^* = 0.96, \quad p_F^* = 0.54, \quad q_I^* = 0.64, \quad q_F^* = 0.36.$$

The optimal inspection probability has hardly changed, whereas the optimal violation strategy is quite different. To answer the last question posed at the beginning we reduce the inspection effort slightly:

$$k = 1.3, \quad 1 - \beta_I = 0.5, \quad 1 - \beta_F = 0.9,$$

and obtain

$$1 - \beta^* = 0.42, \quad p_I^* = 0.84, \quad p_F^* = 0.46, \quad q_I^* = 0.64, \quad q_F^* = 0.36.$$

The violation strategy is unaffected.

Finally, note that our model precludes consideration of deterrence. The option of *legal behavior* is not allowed to enter the Ruritanian Minister's evil mind. This flaw will be dealt with first in Chapter 6, then more systematically in Chapter 9.

1.3 About This Book

In each of Chapters 2 through 8 one or two major problem areas, together with various specific examples, are selected to illustrate applications of verification theory. These chapters are relatively self-contained and may be read independently, although Chapters 3, 4 and 8 deal extensively with statistical decision theory and are best read in that order. Additional mathematical bases for the material treated may be found in Avenhaus (1986) and in many publications on the subject of inspection and verification in nuclear materials control, most of which have appeared in the proceedings of conferences sponsored by the International Atomic Energy Agency (IAEA), the European Safeguards Research and Development Association (ESARDA) and the Institute of Nuclear Materials Management (INMM), as well as in the Journal of the INMM and in the ESARDA Bulletin.

The evident bias of the quantitative verification literature toward nuclear safeguards is a consequence of the existence of a stringent, well-established verification regime for the peaceful use of nuclear energy. Nuclear non-proliferation by no means exhausts the field of application, however. We have attempted to demonstrate this with several examples chosen from fields outside nuclear safeguards. Some of these may appear a little artificial or speculative, given the lack of clearly defined boundary conditions and verification goals, but they are really meant to encourage those involved in the respective fields of application to develop their own, more realistic quantitative models along our proposed lines.

Chapter 9 is rather special, as it is intended both as a deepening of the material preceding it as well as an overview for the entire book. Whilst the first eight chapters deal in effect with the quantification and optimization of assurance achieved through verification, and discuss more or less concrete inspection problems, the theme of the ninth chapter is the quantification of deterrence. Quite general models are formulated and verification theory is treated on a more abstract, sophisticated level, with particular emphasis on the most fundamental aspect of all: *inducing the inspectee to behave legally.*

Perhaps the best way to read the book is as follows:

(i) Read the first chapter. You've presumably just done this.
(ii) Read any subset of Chapters 2 through 8. The systematic reader might recall that, included in the set of subsets of a set, is the set itself. The impatient reader anxious to get to the theoretical meat of Chapter 9 might choose the empty subset.
(iii) Read Chapter 9.
(iv) Repeat step (ii), choosing, of course, a non-empty subset.

Finally, it should be stressed that, beyond some simple calculus and a little familiarity with error propagation, no sophisticated mathematical or statistical knowledge is presupposed on the part of the reader. Where things get slightly advanced, as for example in Sections 5.2 and 6.2.1, the explanations become more detailed. If she managed to survive the trip to Ruritania with all her luggage, the reader should have no difficulty with the rest of this book.