ASIA & THE PACIFIC POLICY STUDIES
Australian National University
Australian Government
Department of Foreign Affairs and Trade
Asia & the Pacific Policy Studies, vol. 2, no. 3, pp. 452–466
doi: 10.1002/app5.100

Original Article

Coping with Nasty Surprises: Improving Risk Management in the Public Sector Using Simplified Bayesian Methods

Mark Matthews and Tom Kompas*

Abstract

Bayesian methods are particularly useful to informing decisions when information is sparse and ambiguous, but decisions involving risks must still be made in a timely manner. Given the utility of these approaches to public policy, this article considers the case for refreshing the general practice of risk management in governance by using a simplified Bayesian approach based on using raw data expressed as 'natural frequencies'. This simplified Bayesian approach, which benefits from the technical advances made in signal processing and machine learning, is suitable for use by non-specialists, and focuses attention on the incidence and potential implications of false positives and false negatives in the diagnostic tests used to manage risk. The article concludes by showing how graphical plots of the incidence of true positives relative to false positives in test results can be used to assess diagnostic capabilities in an organisation—and also inform strategies for capability improvement.

Key words: risk, management, Bayesian inference, public sector

* Matthews: Australian Centre for Biosecurity and Environmental Economics, Crawford School of Public Policy, The Australian National University, Canberra, Australian Capital Territory 0200, Australia; Kompas: Australian Centre for Biosecurity and Environmental Economics, Crawford School of Public Policy, The Australian National University, Canberra, Australian Capital Territory 0200, Australia, and Centre of Excellence for Biosecurity Risk Analysis, The University of Melbourne, Victoria 3010, Australia. Corresponding author: Matthews, email <mark.matthews@me.com>.

1. Introduction

This article considers the case for rethinking how risk is managed in the public sector in general, and in organisations responsible for handling security-related risks in particular. It performs a pragmatic role in drawing the attention of practitioners to practical concerns over the current effectiveness of risk management methods while also performing a 'translational' role by raising awareness of the relevance of Bayesian inference to these practical challenges and by explaining how these methods can be used by non-specialists as part of a reformed risk management agenda.

This question over the case for refreshing approaches to risk is posed because of the confluence of three factors that have, arguably, created a situation in which approaches to risk are increasingly pervasive—yet can be ineffective. As Michael Power has observed:

> Risk talk and risk management practices, rather like auditing in the 1990s, embody the fundamentally contradictory nature of organisational and political life. On the one hand there is a functional and political need to maintain myths of control and manageability, because this is what various interested constituencies and stakeholders seem to demand. Risks must be made auditable and governable. On the other hand,

> there is a consistent stream of failures, scandals and disasters which challenge and threaten organisations, suggesting a world which is out of control and where failure may be endemic, and in which the organisational interdependencies are so intricate that no single locus of control has a grasp of them.
> [Power 2004, p. 10]

These problems were especially evident in the actions addressed by Australia's recent Royal Commission into the implementation of the home insulation scheme (Hanger 2014). The Royal Commission found that the Australian Government's management of risk had been seriously deficient in a number of interrelated respects. These can be characterised as a failure to engage with risk as an integral aspect of the program in preference to treating it as an impediment to rapid policy delivery and with no overall responsibility for managing risk being defined and acted upon. Risk management was treated, in effect, as a compliance ritual rather than as a serious approach to identifying and reducing risks.

Our focus and criticisms of current practice are not directed at existing technical work in risk analysis and assessment—approaches that are very useful in shedding light on, and informing decisions about, complex and challenging risk-related problems. Rather, we address the more general issue highlighted by Michael Power, and framed in the broad context of how 'managerialist' approaches in the public sector in combination with the dominant ethos of evidence-based policy-making and an increasingly pervasive emphasis on risk management, framed as risk avoidance, create a situation in which risk management involves voluminous standards and guidelines applied to what governance involves—but surprisingly sparse technical assistance as regards actually identifying, assessing and dealing with risks as a core aspect of governments' role as uncertainty and risk managers of 'last resort'. Indeed, the proposed solution to this general challenge of refreshing risk management relies upon finding practical ways of transforming the Bayesian approaches already being used by technical specialists for use by non-specialists.

At present, while the importance of adaptive learning based on keeping options open in an uncertain and changing world is a well-established managerial principle (see Klein and Meckling 1958 for an influential contribution), after nearly 60 years there are still pleas for governance to transition to a more adaptive learning-based and 'experimentalist' mode (Sabel and Zeitlin 2012). Although the framing may be different (Klein and Meckling were concerned with weapons system development decision-making, and Sabel and Zeitlin stress a multilevel governance context), the underlying principles, as reflected in governance, have been constant. The challenge is to find practical ways for the public sector to operate in an adaptive learning-based and generally experimentalist manner while still being able to demonstrate value for money.

This article seeks to respond to the challenge set out by this prior work that stresses shortcomings in risk management by considering practical responses that government departments and agencies can start to develop to improve their risk management effectiveness.

2. Appraisal of the Current Situation

Contemporary notions of good governance reflect the confluence of a three policy narratives that describe how efficiency and effectiveness can be achieved in the public sector.

First, the 'new public management' ethos characterised by the privatisation of certain public services and a strong emphasis on the use of targets and performance measures to drive performance and demonstrate transparency and accountability (Hood 1991).

Second, the concept of evidence-based policy-making that places a high priority on the collection and analysis of data as a basis for making policy decisions and makes explicit claims about avoiding 'ideological' issues (Solesbury 2001).

Third, the strong and pervasive role of formal process compliance-based approaches to risk management reflected in various standards and guidelines, and especially in ISO: 31000: 2009.

In combination, this mix encourages decision-makers in the public sector to approach risk as:

- something that should be avoided and/or displaced onto others (Power 2004)—an aspect that can make it difficult to integrate risk into effective strategic plans
- a compliance-based impediment to policy delivery rather than an integral component of what delivering policy in a prudent manner actually involves doing (Hanger 2014)
- being treated as a failure to achieve clearly defined objectives with risk being defined as 'uncertainty over objectives' (ISO 31000: 2009)
- a matter of maintaining 'risk registers' based on the assumed likelihood of occurrence and severity of potential impact—but with no overall measure of resulting risk exposure
- a tendency to overlook the implications, for decision-making, stemming from treating substantive uncertainty (that is, incalculable risk) as if it is calculable risk, and as a result missing opportunities to learn-by-doing in coping with the inherent uncertainties in governing (Matthews, 2015).

The result is a situation in which the management of real, and often unavoidable, uncertainties and risks by officials is hampered by a reluctance to speculate and conjecture (as that goes beyond the available 'evidence'). Indeed, the concept of evidence used in the public sector tends to err towards a preference for 'facts', and especially simple 'killer' facts used to justify a desired and media-friendly intervention (Stevens 2011), rather than a more scientific approach based on the empirical testing of hypotheses.

This aversion to speculation and conjectures is problematic, as Andrew Stirling has observed:

> Since contemplating the unknown necessarily requires imaginings beyond the available evidence, it is treated as unscientific in conventional risk regulation. What is truly unscientific, however, is this effective denial of the unknown. [Stirling 2009]

In the real world of risk management, adversaries and natural phenomenon are unpredictable. However, this does not mean that resulting risks cannot be considered and prepared for—including situations in which an adversary may prepare to act to exploit, and amplify, the damage caused by an unpredictable natural crisis or crises. In short, effective risk management is best approached as an effort to reduce the potential for nasty surprises by combining creativity and conjectures about what *could happen*, with a recognition that aiming simply to comply with prevailing risk management standards and guidelines can, in some circumstances, amplify rather than reduce the potential for nasty surprises (by engendering complacency and passivity once rules are complied with).

While the current emphasis within government is on basing decisions on robust evidence, this is not a basis for effective risk management. It is wise, therefore, to be clear about the differences between concepts of 'proof' using evidence and the tasks for intelligence, which operate under far more ambiguous, time-constrained and fluid conditions. As ex-CIA officer Bruce Berkowitz comments on the distinction between intelligence work and detective work:

> Detective work and intelligence collection may resemble each other, but they are really completely different. Detectives aim at meeting a specific legal standard—'probable cause', for example, or 'beyond a reasonable doubt' or 'preponderance of evidence'. It depends on whether you want to start an investigation, put a suspect in jail or win a civil suit. Intelligence, on the other hand, rarely tries to prove anything; its main purpose is to inform officials and military commanders. The clock runs differently for detectives and intelligence analysts, too. Intelligence analysts—one hopes—go to work before a crisis; detectives usually go to work after a crime. Law enforcement agencies take their time and doggedly pursue as many leads as they can. Intelligence analysts usually operate against the clock. There is a critical point in time where officials have to 'go with what they've got', ambiguous or not. But the biggest difference—important in all the current controversies—is that intelligence agencies have to deal with opponents who take

> countermeasures. Indeed, usually the longer one collects information against a target, the better the target becomes at evasion. So do other potential targets, who are free to watch.
> [Berkowitz 2003]

In short, if we base risk management exclusively on evidence, and if we compound this by relying on fixed performance targets that limit the capacity to learn and adapt to risk, then the outcome is likely to be amplified rather than reduced levels of risk.

3. Current Approaches to Risk and Their Consequences

There are presently two main ways in which risk is defined:

- as the likelihood that something may happen and the magnitude/severity of the consequences if it does happen;
- as the effect that uncertainty has on the achievement of objectives (ISO 31000: 2009).

It is common for risk management in the public sector to be based on classifications using 'risk matrices' that relate likelihood to consequences, and use scoring techniques to aggregate across measures of probability and consequences.

The use of risk matrices, while useful in some settings, can also increase the potential for nasty surprises, and hence amplify risk.

There are five reasons for this concern. First, the matrices specifically downplay low probability and high consequence outcomes. The 'red zones', or the areas that score highest, are in the high probability and consequence categories. While this is fine in itself, it generally shits attention from the outcomes that are worthy of considerable attention.

Second, the matrices always result in 'range compression' (Cox 2008). In one category or box, for example, the designated range in probability measure may range from, say, 0 to 20 per cent. The problem here is that for many especially high consequence outcomes, a change in probability from 10 to 15 per cent of a given outcome can be crucial. But this distinction is simply buried in the given range or box being considered.

Third, the ranges themselves do not always map out in symmetric boxes (for example, 20 per cent blocks), and this can cause confusion over range intervals and what is being measured.

Fourth, a lack of a 'common language' often causes misunderstanding. Categories in the matrix designated as 'catastrophic' or 'almost certain' can mean very different things to those who do risk assessments.

Finally, risk matrices totally obscure problems with 'false negatives' and 'false positives' in security and risk measures, as discussed later, and can never account for 'jumps' in probability assessments or states of nature that are common with nasty surprises. This latter point is essential. Probability measures of potentially severe outcomes cannot only change or 'drift' from one box in the matrix to another over time, but take discrete jumps. Accounting for these jumps and militating against them is especially important for high consequence events.

Risk matrices aside, it is worth stressing that current risk management standards and guidelines place a strong emphasis on the risks faced in achieving stated objectives—with these objectives treated as a given (in effect as an independent variable). While there is a useful emphasis on continuous improvement in risk management practices, there is not a strong emphasis on the ways in which the decisions made in setting objectives will influence the nature and extent of the risks faced.

This point can be summarised by inverting the ISO 31000: 2009 definition of risk as: *the effect of objectives on uncertainty*. Clearly, the more ambitious the objectives, the greater the uncertainty faced in achieving these objectives. Engineers are familiar with the ways in which system complexity increases the likelihood of system failure due to complex interactions and cascading failure modes. This is why engineers prefer to simplify designs in order to reduce the likelihood of failures (and build in redundant/duplicated systems where failures are most likely).

This relationship between objectives, uncertainty and risk forms the basis of well-developed design, development and demonstration in engineering systems (most evident in NASA and U.S. Department of Defense structured program management methods for complex engineering systems). Those methods explicitly focus on attempting to balance the uncertainty (hence risk) associated with attempts to achieve technical objectives with the benefits of actually achieving those technical objectives, using a perspective laid out in Klein and Meckling (1958). This results in a system design tradeoff between aiming for the most technically demanding and potentially most useful objectives ('stretch targets') and the uncertainties and risks faced in attempting to meet those objectives.

In many circumstances (especially when major wars with technologically sophisticated adversaries are not happening), this tradeoff results in less ambitious objectives being set in order to reduce the likelihood of failure to acceptable levels. Scrutiny of this tradeoff between objectives and uncertainty/risk also results in efforts to develop more effective innovation pathways that minimise uncertainty and risk—notably in the readoption of NASA Apollo program-style incremental modular approaches to system evolution (known as rapid spiral development (RSD)).

RSD involves the deliberate prioritisation of modular systems designs that allow incremental advances and testing of discrete system components while holding other module designs constant. The best example is the way in which the Apollo space program tested discrete system components and their use in several successive missions—an explicit risk management technique. The phasing out of the space shuttle in preference for a return to the RSD approach, in the form of the Orion program currently being developed by NASA, reflects recognition that a modular system design allows functionality to evolve in each mission, and for new technologies and operational procedures to be tested and adopted far more easily and cheaply than when using a fixed system such as the non-modular space shuttle (which was forced to operate with very outdated computer systems because it was so costly up update subsystems within that rigid system design).

Consequently, the RSD methods have the potential to inform innovation and risk management in the public sector precisely because they reduce the risks faced at a given level of uncertainty over objectives.

It is also important to stress that while the national security community has key overarching objectives, such as reducing terrorist threats, the core capabilities (preparedness/readiness and rapid response to hard to predict acts) reflect objectives that are set by adversaries—not as part of 'enterprise risk management' frameworks oriented to well-defined agreed and locked-in objectives. Arguably, this aspect of national security practice limits the utility of standards such as ISO 31000: 2009.

4. Using a Simple Diagnostic Ratio Based on the Potential for Surprise

Writing shortly after the end of the Second World War, Claude Shannon distinguished between information and uncertainty (defined as entropy) and expressed the value of new information that might be received in terms of the assumed likelihood of an event happening and being observed. In that framework, which has been incredibly useful in information technology, the less likely an event is assumed to be, the greater the information gain *if* it is observed (Shannon 1948).[1]

Shannon's use of the concept of entropy in the Second World War had a very specific national security objective: the need to be able to calculate the minimum volume of information required to encrypt a message given the statistical frequencies with which different letters are expected to appear for linguistic reasons. Shannon entropy is therefore an expression of statistical uncertainty based only on available information (or noise) rather than being treated as a sample of additional but

1. See Pierce (1961) for a non-technical explanation of Shannon's work.

unobserved data (as in frequentalist/sampling theory-based statistics).

The analytical value of Shannon entropy (as this approach is now known) lies in maintaining and using this distinction between information and uncertainty. In information theory, this distinction allows for tractable calculations of highly complex things, especially error identification and correction and signal to noise ratios, and has provided an analytical framework that has assisted a range of technological advances to be made (including noise and signal error correction in wi-fi).

In this context, Shannon's principle that the less likely an event is assumed to be the greater the information gain *if* it is observed (a surprise-based notion) can be framed explicitly in terms of the potential for surprise in a public policy context using the following ratio:

$$\text{Risk amplification} = \frac{\text{Achieved potential for surprise}}{\text{Unavoidable potential for surprise}}$$

This ratio provides a conceptually simple means of framing risk management more clearly in relation to the potential for nasty surprises—while also having the advantage of opening up an avenue for using measures of information entropy as part of the risk management toolkit.

The risk amplification ratio reflects the reality that governments face a range of complex and unavoidable factors that can surprise them, and that the real objective of risk management is to minimise this potential for nasty surprises (rather than simply comply with risk management standards and guidelines), in particular by seeking to minimise the extent to which there are missed opportunities to learn from practical experience in handling risks and by failing to spot 'weak signals' of potential large and unexpected shifts in circumstances.

This framework based on the potential for surprise (and implementable as entropy measures) provides a basis for using Bayesian tests of competing hypotheses as a risk management method. This is because it emphasises the utility of these hypothesis tests as a means of reducing the amplification of risk. Prudent risk management involves working with a range of hypotheses that, in combination, reduce the ratio of the achieved potential for surprise relative to the unavoidable potential for surprise. The result is, of course, exactly the sort of 'bird's-eye view' of risk exposure that is so lacking in conventional risk register frameworks used in the public sector.

Significantly, risk is defined in a manner aligned with the most recent incarnation of the international risk management standard ISO31000: 2009 as 'uncertainty over objectives'—yet with the major advantage of fostering a constructive focus in the extent risk management practices and procedures actually reduce the potential for nasty surprises.

Consequently, this risk amplification ratio provides a clear high-level measure of both current capability challenges in governance and a basis for planning future improvements in capability and assessing the progress made. The simplicity of a bird's-eye view metric of this type is important because one can easily get lost in the complexity and detail of real governance processes and procedures—approaches that seem to amplify rather than simplify complexity.

Governments should aim to minimise the extent to which the assumptions they hold over the range of likelihoods and consequences of wanted and unwanted events (that is, risks *and* opportunities) are distorted by failures to use all available information on those likelihoods and consequences.

Three examples of risk management failures can be used to illustrate how risk can be amplified by choices made over how to handle uncertainties and risks. First, as noted in the introduction, Australia's recent Royal Commission into the implementation of the home insulation scheme highlighted the way in which a top-level political priority on 'delivery' resulted in risk management being sidelined and treated as a 'speed bump' to delivery, making it hard for officials to adopt a more prudent approach to risks (Hanger 2014). This was exacerbated by the prevalent tendency in government to treat risk management as a compliance ritual to be got out of the way as

easily as possible, and where possible even outsourced to consultants for convenience. This 'sidelined' stance meant that prior information on relevant risks from States and New Zealand was effectively ignored. The Royal Commission's investigation revealed both specific departmental capability shortcomings but also pointed to more general systemic limitations in risk management in the loosely federal Australian system of government: (i) risk as an inconvenience in delivering policy and (ii) a 'tick box' (do and forget) mentality rather than the basis for learning and adaptation, and (iii) poor flows of relevant information across administrative boundaries. In short, this tendency to treat risk management as a compliance ritual is *itself* an amplifier of risk.

Second, NASA's Challenger space shuttle disaster in 1986 demonstrates how technical choices made over risk management methodologies can amplify risks. In the lead-up to the disaster, there were two alternative risk assessments for the space shuttle system (McGrane 2011).[2] Odds of a 1 in 100,000 risk of catastrophic system failure calculated by NASA differed markedly from the findings from a 1983 U.S. Air Force funded review by Teledyne Energy Systems that used Bayesian methods to put the risk of catastrophic system failure much higher (and dangerously so) at odds of 1 in 35. Teledyne had examined failure rates in similar solid fuel rocket systems (as used in Poseidon submarine missiles, and Minuteman intercontinental ballistic missiles) and used these estimates as prior risk factors in the Bayesian analysis (with a prior of 32 confirmed failures out of 1,902 rocket launches) augmented by subjective probabilities and based on lessons from real operating experience. Perhaps due to the Bayesian impact on cryptography in the Second World War, the Pentagon has always been more receptive to Bayesian concepts than NASA. Worryingly, NASA instructed the company it had hired to study shuttle risks to ignore such 'prior' data—even though the rockets were essentially identical. This was partly because NASA's risk management methodology worked on system-specific technical engineering safety margins that sought to eliminate risk via system redundancy rather than use probabilistic risk assessments that have stronger implications for operational decision-making able to respond to unusual events (such as cold weather)—a narrow 'hard' data-driven approach that, in fact, amplified risk.

Third, there is the case of the Hoover Dam. This was designed on the basis of what later turned out to be statistical data on rainfall and consequent river levels from an unusually dry period. This leaves it at risk of collapse unless water is released to pre-empt a rapid rise in the dam level caused by the spring snow melt, in turn the combined consequence of rainfall and snow depositions earlier in the year and the rate at which the snow mass melts due to rising temperatures. As a result, the dam came close to failure in 1983 partly because this pre-emption was not carried out—a decision influenced by assumptions over expected deluge likelihoods (the likelihoods of actual water inflows threatening the dam are greater than the *assumed* likelihoods). These likelihoods are best updated and used to drive operational decision-making rather than sticking to the rule book (based on non-updated likelihoods). The Hoover Dam illustrates the way in which assumed statistical distributions used in design specifications, coupled with risk management solely as compliance (especially when used to define regulatory frameworks), can also amplify risk. The global financial crisis is another example of how regulatory stances based on assumed or unrepresentative statistical distributions can amplify risks in this way.

One key lesson from these examples of risk amplification is that the incidence of false positives and false negatives in diagnostic tests used in risk management can be and an important source of the amplification of risk. In each of the examples summarised above, there was a false negative conclusion to the test for system failure risk. In the case of the space shuttle, the false negative was caused by methodological restrictions imposed by NASA on itself that constrained risk assessment as regards prior data on failures in similar

2. This summary also draws on a range of internal NASA documents bearing upon on risk management shortcomings.

systems and avoided a probabilistic approach to risk management in preference for engineered safety margins (a methodological choice that impacted upon decision-making over risks). In the case of the Hoover Dam, design tolerances and operational guidelines that impact on risk were based on biased statistical data on rainfall patterns. In the case of the Australian home insulation scheme, false negatives arose because risk management was treated as a compliance ritual and an impediment to rapid program delivery (rather than being placed at the centre of the design and delivery of the intervention).

These false negatives amplified the potential for nasty surprises above unavoidable levels. In other words, risk was amplified because organisations made choices over risk management that increased the likelihood of false negatives for tests of potential system failure. In addition, as the discussion of the limitations to some forms of risk matrix has illustrated, this approach can also amplify risks by increasing the potential for nasty surprises.

Given the importance of this 'amplification' aspect of risk management, an analytical means of dealing with the challenge of reducing false negatives (and false positives) in tests relating to risk is provided in the following section. The key enabler of this approach is to frame risk management as binary (true or false) hypothesis tests and to break down complex sets of interrelationships into these binary links.

5. Using Bayesian Signal Processing and Machine Learning Concepts to Articulate the Potential for Surprise

Bayesian probability and the associated statistical methods differ in marked ways from the alternative (and more established) *sampling theory* approach (sometimes referred to as frequentalist, classical or orthodox probability).[3] In the *sampling theory* definition, probability is treated as the long-run relative frequency of the *observed* occurrence of an event. The sample set can be either a sequence of events through time or a set of identically prepared systems (Loredo 1990).

In contrast, Bayesians treat probability (in effect) as the relative *plausibility* of propositions when incomplete knowledge does not allow us to establish truth or falsehood in an absolute sense. This methodological distinction is useful because a Bayesian approach aligns better with learning processes (it is based on updating estimates of the odds that hypotheses are true whenever new data are obtained). Bayesian methods are also well suited to handling the potential for surprise for the same reason. Indeed, information theory itself is derived from Bayesian principles, hence Shannon's emphasis on valuing information in inverse proportion to the estimated likelihood of receiving that information (the definition of surprise that defines the powerful concept of Shannon entropy).

However, a major problem is that, as Ferson comments, Bayesian approaches are rather like snowflakes in the sense that each one is unique (Ferson 2005). This heterogeneity makes it hard for non-specialists to adopt these methods in a public policy context. The resulting dilemma is that while Bayesian inference is, *in principle*, of great relevance to risk management in the public sector (and more general 'risk-aware' applications), the current reliance on complex and technically sophisticated bespoke applications greatly limits the ability of the public sector to use these methods in a day-to-day manner.

Our suggested response to this problem is to develop a simplified and standardised Bayesian expression of the familiar policy learning cycle specifically designed for use by non-specialists. This is a long-term project to be carried out, wherever possible, in partnership with public sector departments and agencies.

The following diagram explains the basic (and very simple) principle behind Bayesian analysis, namely that we are able to update the assumed odds of something happening when new information is obtained. The new information may either confirm that the initially assumed odds should be retained or may lead us to revise these odds.

3. Two useful introductory sources on Bayesian inference can be found in Jaynes (1984) and Fenton and Neil (2013).

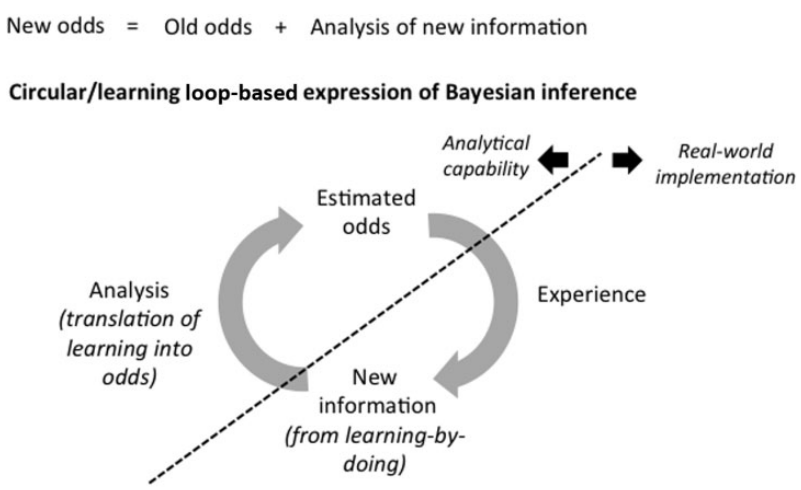
For the purposes of relating Bayesian inference to the policy cycle, the simple equation expressing new odds and a product of the old odds plus the analysis of new information is reframed as a circular learning process. This is illustrated in Figure 1. Estimated odds, via experience, generate new information that when analysed allows estimated odds to be updated. This learning loop combines a real-world implementation phase (experiments in effect) with an analytical phase. Everything that government does is, in effect, in the implementation phase and consequently an experiment (either explicitly or implicitly). However, implementation/experimentation activities may not necessarily involve the new information being identified, collated and analysed. If the latter does not happen, then the odds cannot be updated, and in effect there has been a missed opportunity to learn (and manage risk in particular).

From this perspective, governments' monitoring and evaluation (M&E) activities will have the greatest utility when the information obtained as a result of experience in implementing a policy intervention is related back to an initial (uncertainty and risk based) assumption of the odds of success assumed for the intervention. If M&E measures are not based on an explicit recognition of uncertainty and risk (that is, the odds of success are not made explicit), then it is unlikely that useful learning will be captured *and used* even if useful learning takes place. This is because uncertainty and risk are marginalised rather than centralised in the analysis.

Figure 2 contains the basic analytical taxonomy used in signal processing and machine learning; this is sometimes (usefully) referred to as a 'confusion matrix' by engineers because it draws out the ways in which binary test results can be wrong, and in combination contradictory, and hence cause confusion. In a machine learning context based on the use of algorithms, this confusion paralyses learning and adaptation. In a policy context, the impact on human judgement can be equally paralysing, or can lead to decisions being made that arbitrarily ignore this confusion. This can lock interventions into problematic developmental pathways if not corrected at later stages.

As the confusion matrix highlights, we should prefer regulatory stances (if expressed as competing hypotheses with binary answers) that maximise the true positive rate and the true negative rate, but that also minimise the false positive and the false negative rates. Whenever there are false positive and false negative test results, the response of the regulatory framework is *itself* a risk to effective policy delivery (actions may be taken that are unnecessary, or actions that should be taken are not taken).

Figure 1 Linear and Circular Expressions of Bayesian Inference

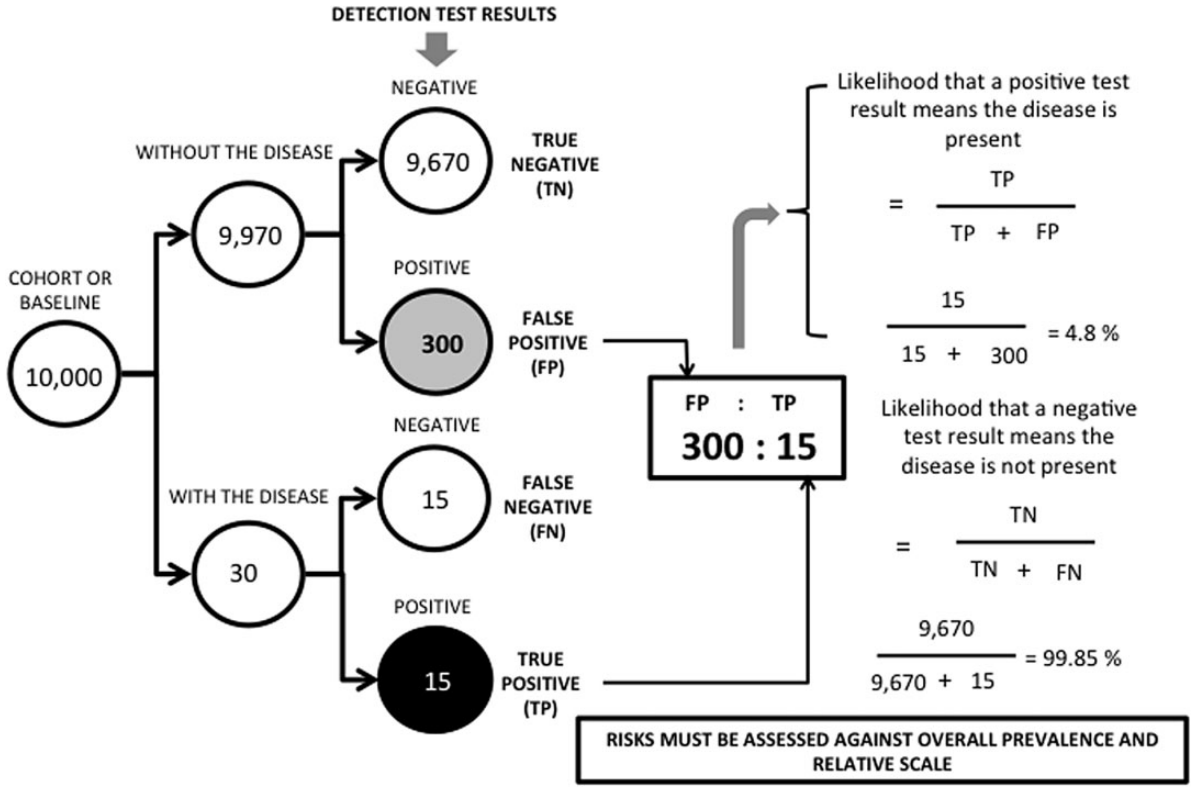


Source: Matthews (forthcoming).

Figure 2 The 'Confusion Matrix' Used in Signal Processing and Machine Learning

Test result	Condition assessment	
	Yes	No
Positive	(a) True positive rate (TP)	(b) False positive rate (FP)
Negative	(c) False negative rate (FN)	(d) True negative rate (TN)

Figure 3 Does a Positive Test Result in Medicine Actually Mean a Condition Is Actually Present?



Source: The authors using data from Gigerenzer (2002).

Figure 3 contains an illustration of the significance of test result errors in a clinical context (using data from Gigerenzer 2002). For many people, this 'natural frequency'-based expression of the situation, which clearly communicates relative scale, is far easier to grasp than the standard Bayesian equation.

In presenting the data in this manner, it is clear that a positive test result (in this case for colorectal cancer) means that there is only a

4.8 per cent likelihood that a particular patient actually has the condition. This is simply because the 3 per cent false positive rate applied to the 9,970 in every 10,000 people who in statistical terms do not have the disease results in 300 cases of false positives relative to 15 true positives. Hence, for an individual patient, one must consider the implications of this ratio of 300 false positives against 15 true positives (the odds from which favour a particular test result being a false positive). This highlights the way in which the overall prevalence of a disease in the population, combined with the rates of true and false positives (and true and false negatives) in test results, generates this gap between a naïve interpretation of a particular test result and a more thoughtful and evidence-based interpretation.

Figure 4 contains the more conventional Bayesian expression of this same situation. Unless one is highly familiar with conditional probability (which many people working in government and in stakeholder organisations are not), then this way of calculating test sensitivity is hard to grasp. This unnecessary complexity is created by avoiding the use of natural frequencies in preference for reliance on the mathematics of probability. As Figure 5 demonstrates, the use of natural frequencies makes Bayes rule far easier to understand—it is simply the sensitivity of the test to the rate of false positives given the prevalence of a condition.

Figure 6 contains an illustration of how this natural frequency approach can be used in a security risk context. In this case, in detecting potential terrorist threats. This illustrates the way in which a very small (0.49 per cent) false positive rate in threat detections can result in a large number of cases treated as threats that are not threats. This diverts scarce resources to dealing with what are believed to be threats that are not in fact threats. These scarce resources would be more usefully directed at the dangerous incidence of false negative threat detections (threats that are not detected).

Figure 4 Conventional Conditional Probability Version of Bayes Rule

$$P(\text{disease}\mid\text{positive}) = \frac{P(\text{disease}) \times p(\text{positive}\mid\text{disease})}{[P(\text{disease}) \times p(\text{positive}\mid\text{disease})] + [p(\text{no disease}) \times p(\text{positive}\mid\text{no disease})]}$$

$$\text{4.8\% (with rounding)} = \frac{0.003 \times 0.5}{[0.003 \times 0.5] + [0.967 \times 0.03]}$$

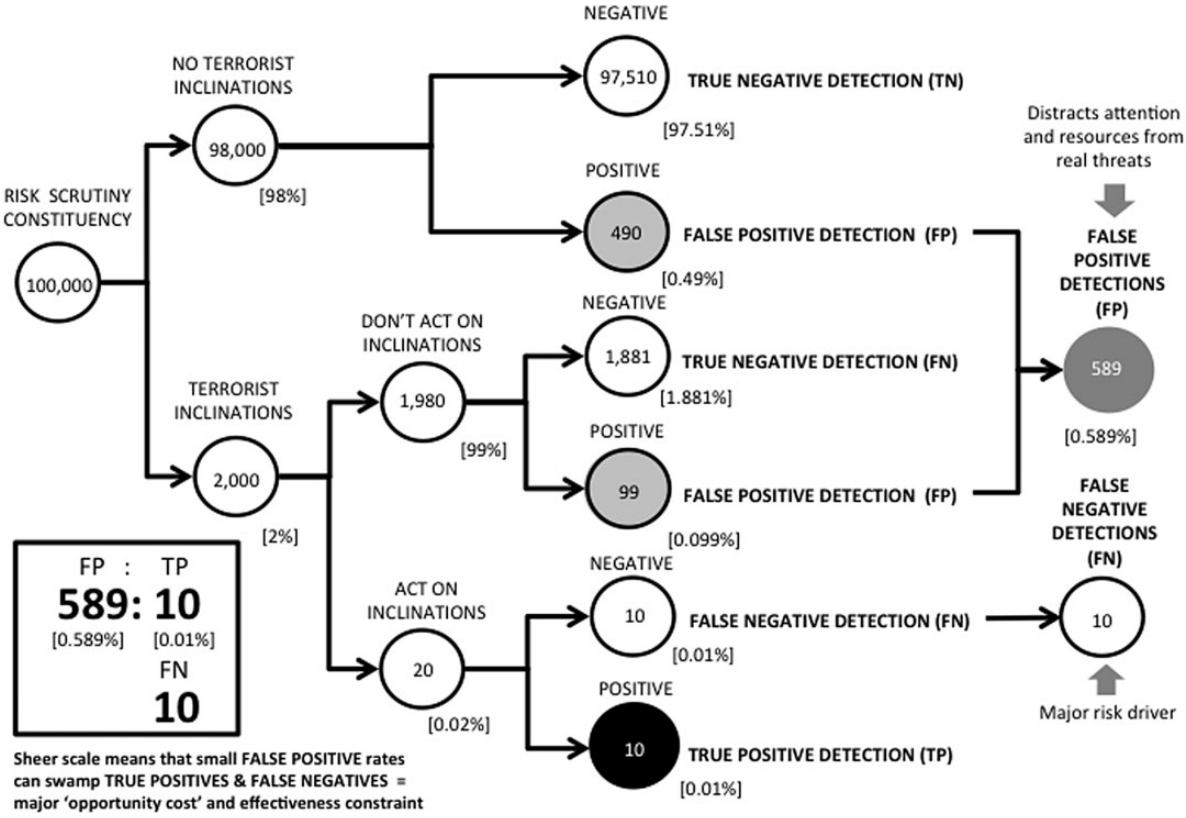
Expressing Bayes rule using conditional probabilities makes it overly complex and hard to grasp. Too convoluted and, luckily, unnecessary.

Figure 5 Natural Frequency Version of Bayes Rule

$$P(\text{disease}\mid\text{positive}) = \text{Likelihood that a positive test result means the disease is present} = \frac{\text{True positive rate}}{\text{True positive rate} + \text{False positive rate}}$$

$$\text{4.8\%} = \frac{15}{15 + 300}$$

Figure 6 Application to Terrorist Epidemiology: Epidemiology = 'Incidence, Distribution and Possibility for Control'



Source: The authors.

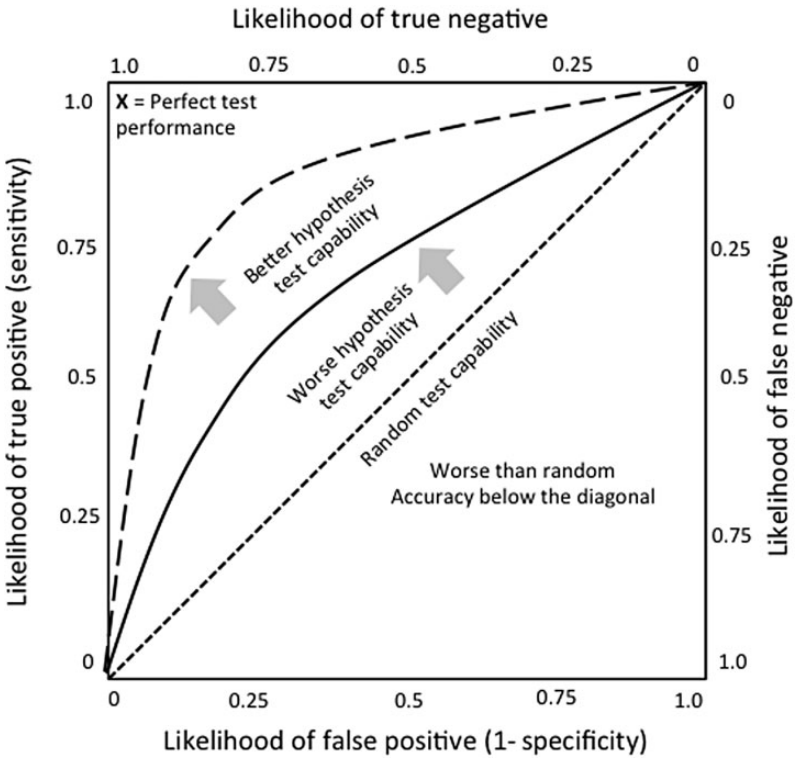
It is easy to see how this use of Bayesian signal processing and machine learning concepts provides a robust and intuitively straightforward basis for assessing aspects of the efficiency and the effectiveness of policy interventions. The approach makes clear where problems caused by test inaccuracies lie, highlights the implications for response decisions and provides a basis for measuring historical changes in diagnostic capability. Crucially, this approach places risk-related concerns centrally in the policy process.

This issue of diagnostic capability is formally expressed in signal processing and machine learning in the following manner (see Figure 7).

For historical reasons, this is referred to as the *receiver operating characteristic* curve (an ROC curve in short). An ROC curve plots the false positive rate (on the X axis) against the true positive rate (on the Y axis) and was originally developed to assess the abilities of radar operators in the Second World War. Some versions, as presented here, also add the true negative and the false negative rates in order to provide a complete diagnostic profile. As such, ROC curves reflect the principles behind the use of randomised control trials (RCTs) in public policy—but in a more generally applicable framework (indeed ROC curves are used in medicine to assess the adequacy of RCT results). For a useful overview of the use of ROC curves in a range of contexts, see Swets et al. (2000).

The best possible performing hypothesis test lies in the top left-hand corner (a test that is 100 per cent sensitive and has a zero false positive rate). Random test results lie on the diagonal (for example, someone guessing the toss result of a coin would expect to eventually end up at the 0.5, 0.5 point in the middle of the diagonal). Test results that are worse than

Figure 7 Measuring Diagnostic Capability Using Signal Processing Methods



random lie below that diagonal (thus providing a particularly useful diagnostic).

In a public policy context, the potential to waste public funds increases the further that capabilities lie from the ideal diagnostic point in the top left-hand corner of the ROC space. Particular test capabilities can be represented as curves in this space: the further above the diagonal and the greater this curvature, the more reliable the hypothesis test is. Shifts in capability over time can be reflected as shifts in these curves.

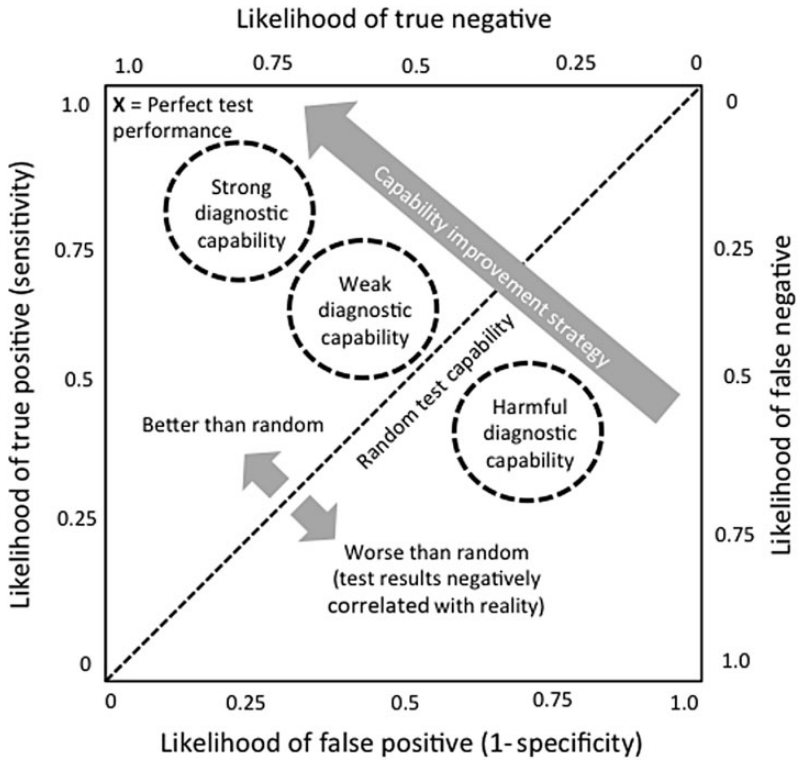
Finally, Figure 8 reinforces the potential utility of this diagnostic framework by indicating possible positions of organisational capability (strong, weak and harmful). In the latter case, test accuracy is worse than random in the sense that test results are negatively correlated with reality. This is not as rare an occurrence in risk management (and indeed public policy in general) as many would assume. This can be caused by cherry-picking 'evidence' to support political aims, weak analytical capacity and other shortcomings that can distort decision-making.

6. Conclusions

This article has drawn attention to the potential that exists to use proven analytical methods widely used in signal processing and machine learning (that are derived from Bayesian inference) as a risk management tool for use in the public sector. The recommended approach, which encourages a 'binary' hypothesis testing-based approach to risk management, focuses attention on the incidence of diagnostic errors (false positives and false negatives) and their implications for risk management. The approach also highlights the ways in which the methodological choices made by an organisation can have the negative unintended consequence of amplifying risks and draws attention to a useful framework for assessing an organisation's diagnostic capability regarding risk management.

July 2015.
The authors would like to acknowledge support for research upon which this article draws provided by the Australian Centre for

Figure 8 Characterising Diagnostic Capability Using Signal Processing Methods



Biosecurity and Environmental Economics (AC BEE).

References

Berkowitz B (2003) The Big Difference between Intelligence and Evidence. Commentary in the *Washington Post*. Re-published by the RAND Corporation.

Cox LL (2008) What's Wrong with Risk Matrices. *Risk Analysis* 28(2), 497–512.

Fenton N, Neil M (2013) *Risk Assessment and Decision Analysis with Bayesian Networks*. CRC Press, Boca Raton.

Ferson S (2005) *Bayesian Methods in Risk Assessment*. Paper prepared for Bureau de Recherches Géologiques et Minières (BRGM), France.

Gigerenzer G (2002) *Reckoning with Risk: Learning to Live with Uncertainty*. Penguin, London.

Hanger I (2014) *Royal Commission into the Home Insulation Scheme*. Attorney-General's Department, Canberra.

Hood C (1991) A Public Management for All Seasons? *Public Administration* 69, 3–19.

Jaynes ET (1984) *Bayesian Methods: General Background*: *An Introductory Tutorial*. Paper presented at the Fourth Annual Workshop on Bayesian/Maximum Entropy Methods, Calgary, August 1984.

Klein B, Meckling W (1958) Application of Operations Research to Development Decisions. *Operations Research* 6(3), 352–63.

Loredo TJ (1990) From Laplace to Supernova SN 1987A: Bayesian Inference in Astrophysics. In: Fougere PF (ed) *Maximum Entropy and Bayesian Methods*, pp. 81–142. Kluwer Academic Publishers, The Netherlands.

Matthews M (2015) *How Better Methods for Coping with Uncertainty and Ambiguity Can Strengthen Government—Civil Society Collaboration, Designing and Implementing Public Policy: Cross-Sectoral Debates*. In: Carey G, Landvogt K, Barraket J (eds) *Studies in Governance and Public Policy Series*, pp. 75–89. Routledge, London.

Matthews M (forthcoming, 2016) *Transformational Public Policy*. Routledge, London.

McGrane SB (2011) *The Theory that Would Not Die*. Yale University Press, New Haven.

Pierce JR (1961) *Symbols, Signals and Noise: The Nature and Process of Communication*. Harper & Brothers, New York.

Power M (2004) *The Risk Management of Everything; Re-Thinking the Politics of Uncertainty*. Demos, London.

Sabel C, Zeitlin J (2012) Experimentalist Governance. In: Levi-Faur D (ed) *The Oxford Handbook of Governance*, pp. 169–183. Oxford University Press, Oxford.

Shannon CE (1948) A Mathematical Theory of Communication. *Bell System Technical Journal* 27(July), 379–423.

Solesbury W (2001) *Evidence Based Policy: Whence It Came and Where It's Going*. ESRC UK Centre for Evidence Based Policy and Practice: Working Paper 1, Queen Mary College, University of London, London.

Stevens A (2011) Telling Policy Stories: An Ethnographic Study of the Use of Evidence in Policy-Making in the UK. *Journal of Social Policy* 40(2), 237–55.

Stirling A (2009) Risk, Uncertainty and Power. *Seminar Magazine* 597, 33–9.

Swets JA, Dawes RM, Monahan J (2000) Psychological Science Can Improve Diagnostic Decisions. *Psychological Science in the Public Interest: A Journal of the American Psychological Society* 1, 1–26002E