CHAPTER

31 Bayesian and Signal Detection Models

Jason S. McCarley *and* Aaron S. Benjamin

Abstract

Cognitive systems, whether human or engineered, must often reason from and act on probabilistic information, and many of their decisions are therefore inescapably uncertain. Such probabilistic decision making is the purview of the two approaches reviewed in this chapter: Bayesian analysis and the theory of signal detection (TSD). Bayes' theorem provides a normative means of updating probabilistic beliefs in light of new data, and modern Bayesian techniques allow decision makers to model joint distributions of large sets of variables. For cases in which a human decision maker must make unaided Bayesian inferences, cognitive psychology has developed and validated simple guidelines for data representation to optimize performance. TSD models the transformation of probabilistic assessments into discrete diagnoses, dissociating the representation of evidence from the decision rules applied to that evidence and establishing normative criteria against which the performance of a cognitive system can be measured. Together, Bayesian and signal detection models offer methods of making, modeling, and assessing judgment and decision making under uncertainty, for both human and engineered agents.

Key Words: decision making, uncertainty, Bayes' theorem, Bayesian networks, receiver operating characteristic (ROC)

The world as we experience it is inherently probabilistic, and every decision maker, human or engineered, contends with uncertainty. Adaptive behavior requires that we update our beliefs as imperfect information accumulates and, eventually, that we render from our imperfect beliefs a choice of action. These two problems are the domain of *Bayes' theorem* and *the theory of signal detection (TSD),* respectively. These methodological tools, widely applicable in cognitive engineering and related domains, provide frameworks for descriptive models of human perception and cognition, and allow us as well to establish normative criteria for decision-making tasks, providing benchmarks against which we can judge a decision maker's performance and guidelines by which we can design artificial agents with optimal decision-making strategies. Bayesian analysis can be applied, furthermore, to draw inferences or make predictions about a human operator's cognitive ability or state (e.g., Jipp, Badreddin, Abkai, & Hesser, 2008; Liang, Lee, & Reyes, 2007) and even to predict the behavior of immensely complex socio-technical systems (e.g., Trucco, Cagno, Ruggeri, & Grande, 2008), while TSD provides an array of performance measures that not only allow us to gauge a decision maker's performance but also to disentangle the agent's underlying ability to discriminate or detect events from higher-level strategic aspects of performance (e.g., Craig, 1987; Meissner & Kassin, 2002; Vickers & Leary, 1983). Here we provide a brief overview of how these theories can be applied to problems in cognitive engineering and human factors. (For fuller tutorial presentations of Bayesian analysis, see Darwiche, 2009; Griffiths,

Kemp, & Tenenbaum, 2008; Koller & Friedman, 2009. For detailed presentations of TSD, see Green & Swets, 1966; MacMillan & Creelman, 2005; Wickens, 2002.)

Bayes' Theorem and Bayesian Modeling

Early efforts at automated reasoning in artificial intelligence aimed for a purely deductive system of inference, using logical rules to reason with certainty from a knowledge base to conclusions, but faltered on an inability to deal with inconsistent, contradictory, or incomplete evidence (Darwiche, 2009). In response, researchers forewent the demand for deductive certainty, pursuing instead a method of probabilistic reasoning more robust, flexible, and considerably more similar to human reasoning than the unattainable deductive ideal (Pearl, 1988). The foundation of this alternative approach was *Bayes' theorem*, a well-known finding in probability theory that provides a normative means of updating probabilistic hypotheses about the world.

Named for the 18th century cleric by whom it was enounced (albeit in a different form than that currently familiar), Bayes' theorem gives the probability that a belief or hypothesis, H, is true given a set of evidence, E. Phrased differently, Bayes' theorem allows us to know how much our confidence in hypothesis H should increase or decrease once we have seen the evidence E. By how much, for example, should a scientist's confidence in her theory increase once an experiment has yielded an outcome consistent with her predictions? By how much should an economic model's estimated risk of a recession increase upon incorporating new data that show a decrease in consumer spending? How much more or less certain should a doctor, or a decision support system, be of a lupus diagnosis after learning that the patient has a family history of the disease but has produced only borderline positive results on a screening test?

In its familiar formulation, Bayes' theorem is:

$$P(H|E) = [P(E|H) \times P(H)] \,/\, P(E).$$

Here, $P(H)$ is the *prior* probability or *base rate* of H, the probability that the hypothesis is true before we have collected any new information, and $P(H|E)$ is the *posterior* probability, the probability of H given evidence E. To ascertain the change in probability as a function of the new evidence E, two other parameters are necessary. $P(E|H)$ is the *likelihood* of E given H, and $P(E)$ is the *prior predictive probability* or *marginal probability* of the evidence E. Thus we begin with a prior, update it in light of new information, and end with the posterior. That posterior can then serve as the prior to be updated in our next round of data gathering.

Treating the elements of Bayes' theorem as random variables allows us to avoid a strong commitment to point values that are usually highly uncertain, and allows us to draw conclusions from the changing shape and location of the posterior distribution rather than simply from a point estimate of its central tendency. Because reasoning in psychology rarely involves point predictions, and because theories of decision making like TSD start with such distributions, this approach is well suited to cognitive modeling. Bayes' theorem tells us that the posterior probability distribution, the belief that we end with after taking into account new evidence, is a joint function of our pre-existing beliefs, as embedded in the prior, and the strength of the new evidence. The higher our pre-existing confidence in the hypothesis under consideration, the higher our posterior confidence. Conversely, our posterior confidence is inversely proportional to the marginal probability of the data themselves. If the pattern of evidence is highly likely whether or not H is true, then the observation of that evidence will do little to influence our posterior beliefs. In the extreme case that $P(E) = 1$, observing E has no influence at all on our beliefs, and $P(H|E) = P(H)$.

As Equation 1 indicates, to calculate a posterior probability requires three pieces of information. Where do these come from? For some purposes, happily, the prior predictive probability $P(E)$ can be treated as a normalizing constant and ignored. In comparing the relative posterior probabilities for a pair of hypotheses H_1 and H_2, for example, we can calculate a *likelihood ratio*, $(P(E|H_1) \times P(H_1))/(P(E|H_2) \times P(H_2))$, dividing out $P(E)$. Establishing the prior $P(H)$ can be more problematic. One solution is to begin analysis with an uninformative prior, assuming that $P(H)$ is distributed uniformly across the range of possible (or plausible) values. Doing this, the we assert no a priori belief about the likeliest value of $P(H)$ at the outset of data collection. The posterior probability resulting from the each round of data observation, however, becomes the prior probability for the following round, and our posterior probability distribution gradually approaches truth (Sivia & Skilling, 2006). As an alternative to using an uninformative prior distribution, we may choose a theoretically motivated informative distribution. The advantage to this choice is that, if the assumed prior distribution is roughly correct, our

beliefs as encoded in the posterior distribution will approach truth more quickly than with an uninformative prior. The disadvantage is that, if the assumed prior distribution is wildly off the mark, more data will be necessary to correct for the false assumptions encoded in the prior, and our beliefs will approach truth more slowly. Interestingly, empirical data suggest that human decision makers' implicit assumptions about the form of the prior distribution, as evidenced by explicit judgments of posterior probabilities, are highly accurate across a range of very different contexts (Griffiths & Tenenbaum, 2006).

The likelihood function $P(E|H)$, finally, is typically given by the statistical model under consideration. To take a common example, consider a simple experiment of 10 coin flips to determine whether or not a coin is fair. Here, the hypothesis under consideration is that the probability Θ of the coin landing heads-up is 0.5, and the datum of interest is the number of heads, x, that results from the 10 flips. The likelihood function, $P(x \mid \Theta = 0.5)$, is therefore given by a binomial distribution with $p = 0.5$. In other cases, unfortunately, the likelihood function may be less obvious. This is especially true when our theory testing or inference requires not just that we weigh a single piece of evidence, but consider a complex pattern of multiple variables. However, such cases often lend themselves readily to a form of graphical Bayesian analysis, as discussed below.

Bayesian Networks

Often, we are interested not just in the relationship between a single datum and hypothesis, but in the contingencies among many pieces of data and multiple measures. More technically, we wish to draw inferences about the joint distribution of multiple random variables. *Bayesian network (BN)* analysis (Pearl, 1988) provides a methodology for approaching such problems. A BN is a graph that, like any, comprises a series of nodes connected by edges. More specifically, a BN is a form of *directed acyclic* graph, where the term *directed* means that connected nodes form ordered pairs, and *acyclic* means that a path leaving any given node cannot lead back to the same node. Because edges are directed, one node within each connected pair can be designated the parent and the other designated the child. Within a BN, each node represents a random variable, and every child node is characterized by a conditional probability distribution specifying the probability of the node's potential values contingent on the value of its parents. The root nodes, those which have no parents, are characterized by marginal (non-conditioned) probability distributions. Within such a representation, every variable, when conditioned on the value of its parents, is independent of every other variable except its descendants, a characteristic known as the *Markov property*. A BN thus provides a graphical representation of the relationships—dependencies and independencies—between variables in a joint probability distribution.

In application, some of the variables comprising a Bayesian network are observable, and others are not. The analyst's goal is to infer the values of the unobservable variables from the observable ones. Imagine, for example, that a cognitive engineer wishes to design a decision-aiding system to judge whether a military pilot is too fatigued to fly a mission (e.g., Rabinowitz, Breitbach, & Warner, 2009). Because the pilot's level of fatigue cannot be observed directly, it must be inferred from observable evidence. Assume that three pieces of evidence will be considered in rendering each judgment: the cumulative amount of sleep the pilot has had in the previous 24 hours, the current time of day (e.g., Gunzelmann, Gross, Gluck, & Dinges, 2009), and a measure of performance on a short battery of psychomotor tests (Dinges & Powell, 1985). Figure 31.1a shows the simplest possible Bayesian network that we might construct from these data. Here, each piece of observable evidence is treated as an independent predictor of the value of the unobservable variable. Such a model is known as a *naive Bayesian network*, and rests on the assumption, evident in the figure, that all observable variables are independent from one another. Despite their simplicity, naive Bayesian nets often perform well. When the assumption of independent evidence variables is strongly violated, however, performance of such models can mislead, overweighting the information provided by correlated variables (Koller & Friedman, 2009). In the current example, for instance, it seems likely that psychomotor performance will be correlated with time of day and with the amount of sleep accumulated in the preceding 24 hours, meaning that a naive Bayes model is probably inappropriate. Figure 31.1b presents an alternative model that might better capture the relationships among the variables under consideration. Here, fatigue is assumed to be contingent on the sleep levels and time of day, while psychomotor performance is assumed to be contingent on fatigue. The structure of a BN will generally be dictated by the modeler's theoretical premises, and, in cases where the appropriate structure is not obvious, the

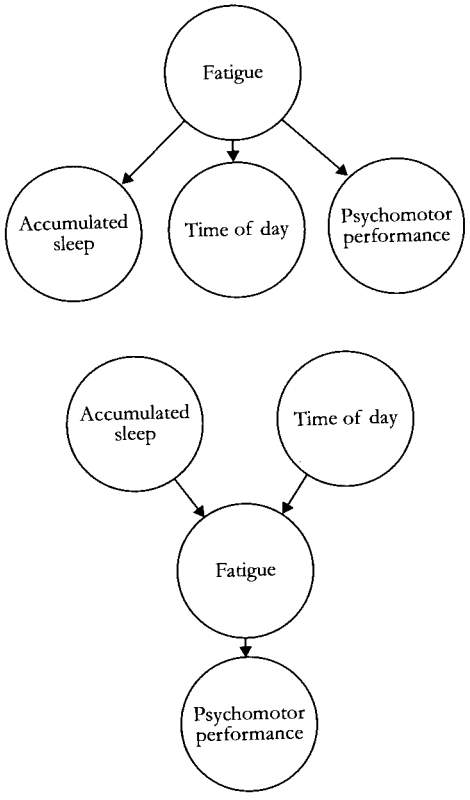

Figure 31.1 A Bayesian network relating sleep, fatigue, and performance.

modeler can design and test the fit of multiple structures to determine which performs best. In the event that theory does not imply any network structure *a priori*, machine learning algorithms can be used to infer structure from data.

Of course, specifying the network structure is only the first step in building a Bayesian network model. Once the structure is in place, it is necessary to attach a probability distribution to each node (marginal distributions for the root nodes, conditional distributions for all child nodes). How are these determined? In some cases, the analyst may be able to solicit estimates from subject matter experts; Renooij (2001) reviews methods for doing this. Note that the Markov property dramatically simplifies the task of establishing conditional probability distributions for the variables in a BN, since each variable need only be conditioned on its parents. Nonetheless, probability estimates for components of complex systems may be difficult to elicit. In other cases, the subject matter experts may not be available. In these instances, machine learning algorithms can again be used, now to infer the needed probability values from pre-existing data or from data collected with the specific purpose of training the model. Returning to the example above, for instance, we might design an experimental study in which pilots come to the lab at different times of day and with different amounts of sleep, complete the psychomotor test battery, then fly a simulated mission. On the basis of his or her simulated flight performance, we could retrospectively classify each pilot as either alert enough or too fatigued to fly, and the data acquired could be used to train the BNs in Figure 31.1 to classify pilot levels based on time of day, accumulated sleep levels, and psychomotor performance. If we were not satisfied that either of the network structures we've specified a priori is appropriate, we might also infer an alternative structure from the data.

Bayesian Reasoning in Humans

Bayesian network analysis provides a methodology for modeling even immensely complex systems of variables. In everyday behavior—in the workplace, the vehicle, and the home—the Bayesian reasoning problems that decision makers face may be much simpler. Generally, though, they must be solved by an unaided human, without the benefit of a BN or other formal support system. How closely do human decision makers approach the Bayesian ideal under such circumstances?

Since Kahneman and Tversky's (1972, 1973) work in the early 1970s on judgment and decision making, conventional thought in cognitive psychology has held that humans are inherently poor Bayesian reasoners. Kahneman and Tversky's data suggested that human decision makers largely ignore event base rates—$P(H)$ in Bayes' theorem—when judging the probability of a hypothesis given a set of data, and research on decision makers in high-consequence domains outside the laboratory appeared to corroborate this conclusion. One study (Casscells, Schoenberger, & Graboys, 1978) presented physicians, hospital staff, and advanced medical students this problem:

If a test to detect a disease whose prevalence is 1/1,000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs?

Assuming the best-case scenario that the test in question has a true positive rate of 100%, the correct answer to this question is approximately 2%; even though the false positive rate is modest (5%), the very low base rate of the disease means that a positive result is far more likely to come from a patient without the disease than from a patient with it. Fewer than 20% of the participants in Casscells

et al.'s study answered correctly, and even later work carefully clarifying the terminology and assumptions of the problem led to a correct response rate of only 36% (Cosmides & Tooby, 1996). Results such as these led many to the conclusion that human decision making is decidedly non-Bayesian.

Recent work, though, has rehabilitated the human decision maker's reputation as a Bayesian reasoner. A variety of behaviors, even if they do not require explicit probabilistic judgments, demonstrate an implicit sensitivity to Bayesian contingencies. Visual attention, for example, appears to prioritize stimuli that are likely to have a large influence on posterior probabilities (Itti & Baldi, 2009), and a Bayesian model of saccade target selection accounts well for the distribution of eye movement amplitudes during reading (Engbert & Krügel, 2010). Perhaps more surprisingly, even explicit predictions of conditional events show an exquisite sensitivity to the form of the prior probability distributions. Griffiths and Tenenbaum (2006) asked participants to predict a variety of real-world phenomena (e.g., the length of a poem, the total grosses of a movie) conditioned on a piece of background information. One item asked, for example, "If you were calling a telephone box office to book tickets and had been on hold for 3 minutes, what you would predict for the total time you would be on hold?" Phenomena about which judgments were made were chosen to reflect prior probability distributions of different forms (Gaussian, power-law, Erlang). Remarkably, participants' judgments implied an awareness of the form of the prior distribution for each of the different types of phenomena.

Moreover, even explicit probabilistic judgments of the type made in Casscells et al.'s (1978) study, described above, can be made to approach the Bayesian norm. Gigerenzer and Hoffrage (1995) and Cosmides and Tooby (1996) speculated that human decision makers' poor performance in earlier studies of Bayesian judgments was not the consequence of an inherent inability to reason in accordance with Bayes' law, but was a side effect of the format in which the problems were presented to study participants. Studies demonstrating poor Bayesian reasoning, these authors noted, generally presented information to participants as probabilities (e.g., a 5% chance of a false positive result) or relative frequencies (e.g., 5% of healthy people tested produce a false positive result). People interacting with the natural environment, however, do not encounter information in these formats. Rather, they experience individual events and accumulate information about event frequencies. Gigerenzer and Hoffrage (1995) and Cosmides and Tooby (1996) thus argued that evolution is more likely to have equipped human decision making to reason with natural frequencies than with probabilities or relative frequencies. Gigerenzer and Hoffrage noted as well that a representation of probabilistic information as natural frequencies dramatically simplifies Bayesian reasoning. Imagine that a person has encountered some number a of events in which evidence E was obtained when hypothesis H was true, and some number b of events in which evidence E was obtained when H was false. The probability $P(H|E)$ can then be estimated easily as

$$P(H|E) = a/(a + b)$$

As predicted by these arguments, experiments have revealed that human decision makers indeed come closer to the Bayesian ideal when reasoning with natural frequencies than when reasoning with probabilities or relative frequencies. For example, Gigerenzer and Hoffrage (1995) compared responses to problems described with probabilities or natural frequencies in the following ways.

PROBABILISTIC DESCRIPTION

- The probability of breast cancer is 1% for women age 40 who participate in routine screening.
- If a woman has breast cancer, the probability is 80% that she will get a positive mammogram.
- If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammogram.
- A woman in this age group had a positive mammography. What is the probability that she actually has breast cancer? ____%

FREQUENTIST DESCRIPTION

- Ten out of every 1000 women age forty who participate in routine screening have breast cancer.
- Eight out of every 10 women with breast cancer will get a positive mammogram.
- Ninety-five out of every 990 women without breast cancer will also get a positive mammogram.
- Here is a new representative sample of women at age 40 who got a positive mammography in routine screening. How many of these women do you expect to actually have breast cancer? ____ out of ____

In a sample of university students, only 16% responded correctly to the probabilistic description above, while 46% responded correctly to the

description of natural frequencies (Gigerenzer & Hoffrage, 1995). In a sample of experienced physicians, even more remarkably, only 8% responded correctly to the probabilistic description, while 46% responded correctly to the natural frequency description (Hoffrage & Gigerenzer, 1998). Cosmides and Tooby (1996) demonstrated a similar advantage for natural frequency descriptions using Casscells et al.'s (1978) medical reasoning problem, and Hoffrage, Lindsey, Hertwig, and Gigerenzer (2000) documented the benefits of natural frequency representations for professional decision makers in AIDS counseling and the law.

The implications of these findings are clear. To encourage Bayesian reasoning, display designers should present probabilistic information using natural frequencies. In fact, not only should displays present natural frequencies, but they should *omit* descriptions of probabilities entirely; probabilities presented redundantly with natural frequencies not only fail to improve reasoning but actually produce poorer judgments than natural frequencies presented alone (Cosmides & Tooby, 1996).

Of course, even under the best of circumstances our probabilistic judgments will (by definition) be uncertain. The assistance of a BN or the representation of data as natural frequencies may improve our chances of a correct judgment, but neither can guarantee a correct judgment. Having made a series of classifications or diagnoses, we can therefore ask ourselves how good—how much better than chance, how close to optimal—those judgments were. This is the purview of TDS.

The Theory of Signal Detection

The Search for Extraterrestrial Intelligence (SETI) Institute, based in Mountain View, CA, is the premiere research organization devoted to seeking evidence for life outside of this planet. A large part of its task is a problem of signal detection—determining whether a particular pattern of radio waves originating from space is an intelligently generated message or merely a product of celestial noise. This is, of course, a complicated and expensive proposition, but is no different fundamentally than the problem faced by a radiologist trying to discern whether a lung x-ray contains a tumor, a security checkpoint screener trying to determine whether a bag contains a weapon, or a Bayesian network trying to classify a pilot as fatigued or alert. Any task faced by an organic or engineered agent in which evidence must be scrutinized in order to determine whether it derives from one of two mutually exclusive and exhaustive sources is a problem of signal detection, and TSD (Green & Swets, 1966) provides a simple framework with which to understand performance and prescribe optimal behavior in such tasks.

TSD achieves its simplicity by avoiding a major substantive problem in any decision task—namely, the problem of how evidence is collected. It is for that reason that the theory is flexible enough to treat problems in both engineering and psychology: It makes no effort at either a theory of microwave antennae or human perception. What it does is characterize the statistical problem of rendering a decision from imperfect evidence: How likely is it that a given evidence sample is consistent with the presence of a signal—a message from aliens, a cancerous tumor, a camouflaged gun, a fatigued pilot—rather than from noise or some other specified alternative? The general approach of TSD is depicted in panel A of Figure 31.2, in which the

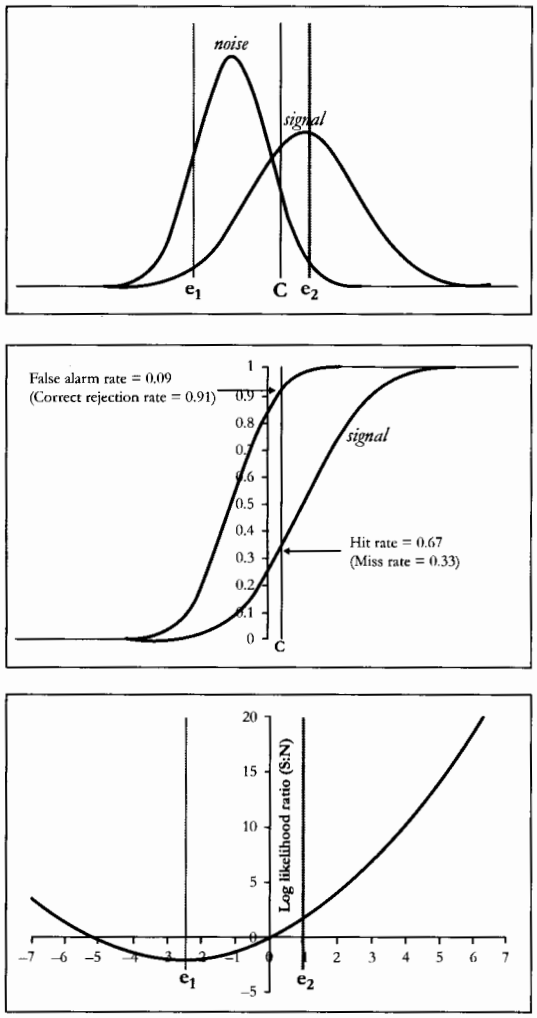

Figure 31.2 Distributions of signal and noise in theory of signal detection (TSD).

abscissa represents the strength of evidence that an evidence sample represents a signal, and the curves represent probability density functions describing the distribution of evidence assuming either that a signal is absent (left curve) or that a signal is present (right curve). The standard signal detection model assumes that the evidence distributions for signal and noise are Gaussian with equal variance. In fact, though, neither the assumption of Gaussian distributions nor the assumption of equal variance is crucial to TSD (Pastore, Crawley, Berens, & Skelly, 2003), and empirically, the assumption of equal variances is often violated (Swets, 1986). In Figure 31.2, the signal distribution is arbitrarily assumed to be of greater variance than the noise distribution. Any experienced level of evidence is represented by a single point on the abscissa (for example, e_1 or e_2), and the corresponding likelihood of that level of evidence occurring under conditions of noise and signal, respectively, are indicated by the height of the probability density functions at those points. In this example, e_1 is more likely to be noise than signal, and e_2 is more likely to be a true signal than noise. The cumulative likelihood of noise and signal sources as a function of the amount of evidence is shown in panel B, and the likelihood ratio is shown in panel C. These values will become important shortly, when we consider how to optimize signal detection performance.

Decision Making

That a psychological or physical experience can be described as a set of likelihoods, each corresponding to a probabilistic hypothesis about the nature of evidence under various conditions, is the first major theoretical tenet of TSD. However, since a decision must eventually be made on the basis of the evidence—to call the president, to order treatment, to detain a traveler, to allow a pilot to fly—a procedure is necessary for translating the continuous variables of evidence and likelihoods into a binary decision. In TSD, this is done by application of a deterministic decision criterion: Evidence greater than a prespecified level is taken to indicate a signal, and evidence below that level is rejected as a non-signal. This mechanism contrasts with other theories of decision making in which evidence is probabilistically related to the decision (Parks, 1966; Schoeffler, 1965) or the criterion is conceptualized to be noisy (Benjamin, Diaz, & Wee, 2009; Mueller & Weidemann, 2009; Wickelgren, 1968).

When an evidence regime is brought together with a decision criterion, we can describe all of the possible outcomes of a decision event by the heights of cumulative density functions, as shown in panel B of Figure 31.2. If a signal event is detected, it is a hit; if it is not detected, it is a miss. When a noise event is mistaken for signal, it is a false alarm; when it is correctly classified as noise, it is a correct rejection. These terms map in a straightforward manner onto the areas to the left and right of the criterion in the mass functions in panel A, and onto the nomenclature used to describe the precision of medical tests: Test sensitivity is hits / (hits + misses), and test specificity is correct rejections / (correct rejections + false alarms).

Clearly, the goal of any decision maker in a signal detection task is to have a high hit rate coupled with a low false-alarm rate. But Figure 31.2 reveals perfect performance to be generally unattainable: When the signal and noise distributions overlap, it is impossible to set a decision criterion that will correctly classify every event. As the distributions overlap less, obviously, then the ability to correctly classify events improves. The separation between distributions is thus described as the decision maker's *sensitivity*. Unless sensitivity is large enough to ensure negligible overlap of the distributions, though, no criterion can guarantee perfect performance. How, then, should a criterion be chosen? A good starting point is to consider the *log likelihood ratio* at the experienced amount of evidence. The log likelihood ratio is shown as a function of evidence in panel C, and provides a rough-and-ready means by which to make a decision. If the log likelihood ratio is greater than 0, then the evidence favors a signal, all other factors being equal. If the log likelihood is less than 0, the evidence suggests the event is more likely to be noise than signal. In the example, e_1 falls below 0 and thus is more likely to be noise, whereas the example event e_2 yields a log likelihood ratio greater than 0 and is more likely to be a signal.

A prudent decision is, however, only partly a function of the evidence sample at hand. Bayes' theorem tells us that the base rates of the underlying conditions should be taken into account when considering what to infer; a wise decision maker will be more reluctant to infer a rare event than a common one. Intelligent extraterrestrial life is pretty hard to come by, cancerous tumors are not, and we should therefore be more inclined to believe evidence of a tumor than evidence of an intelligent extraterrestrial message. In the detection framework, event base rates influence the conservativeness with which the decision maker sets the response criterion.

To be more specific, Bayesian analysis tells us that to decide whether an event is signal or noise, we should scale the likelihood ratio based on the evidence by the ratio of the signal and noise event base rates, yielding a quantity known as the *Bayes' factor*. A criterion at the location on the evidence axis where the Bayes' factor equals 1 maximizes response accuracy.

In many contexts, though, the decision maker's goal is not to maximize response accuracy but to maximize the expected value of his or her judgments. Failing to prevent the outbreak of a virulent disease has serious public health consequences, for example, whereas a mistaken positive diagnosis of the disease may lead to containment efforts that are, by comparison, inexpensive. A public health official may therefore wish to err on the side of over-diagnosing the disease rather than risk even a moderate number of failed diagnoses. More generally, the costs and benefits of the various decision outcomes need to be considered when setting a judicious criterion in any signal detection task. If the cost of a miss is less than the cost of a false alarm, it is wise to adopt a liberal decision criterion. If the cost of a false alarm is greater than the cost of a miss, conversely, a more conservative criterion is optimal.

In practice, other factors influence the placement of a decision criterion, such as the experienced variability of evidence across a range of decisions (Benjamin & Bawa, 2004; Hirshman, 1995). Reviews of how criteria are adjusted are provided by Rotello and MacMillan (2007), and a theoretical proposal for the placement and adjustment of criteria is provided by Treisman (1987). For present purposes, it is enough to see that TSD provides a means by which to disentangle aspects of performance related to decision making (the criterion) and aspects related to discrimination (the ability to accurately discern signal from noise). The following section discusses this further.

Separating Bias from Sensitivity: The Receiver Operating Characteristic

A useful conceptual and analytic tool in TSD analysis is the *receiver operating characteristic (ROC)*. The ROC is a plot of the hit rate against the false alarm rate across all possible criteria, and is an example of a larger class of functions known as state-trace plots (Bamber, 1979). Because criteria serve only to partition the evidence space into binary decisions and do not influence the location and shape of the evidence distributions, the ROC provides a bias-free measure of performance. An ROC for the distributions displayed in Figure 31.2 is shown in Figure 31.3. Several aspects of the plot are revealing. First, performance is related to how far the curve is from the major diagonal, which represents chance levels of discrimination. The most common measures of sensitivity are estimated by the distance from that curve from the major diagonal and the area under that curve. Distances from the diagonal are measured in standard deviation units of the noise distribution (panel A, Figure 31.1) and include d', d_a, and d_e. These differ from one another only in how they account for differential variance of the two distributions. Whether the distributions differ in variance is also evident from the ROC; in this case, the greater mass under the (not shown) minor diagonal indicates that the signal distribution is of greater variance. Measures for area under the curve include A_z and A_z' which differ in how the area is estimated.

Any point on the ROC function reflects a possible criterion. The ROC makes it easy to see why it is absolutely necessary to account for differences in response policy in order to understand discrimination: A conservative decision maker can maintain a very low hit rate and be equivalent in sensitivity to a liberal decision maker with a high hit rate. TSD can be thus be considered a tool used to re-parameterize experimental statistics—hit and false-alarm rates—into the theoretically meaningful variables of sensitivity and response bias.

In the example here, the ROC can be directly inferred from the distributions shown in Figure 31.2. In application, those distributions can be estimated only from empirical performance data in a detection task. Because a single hit rate/false-alarm rate pair is only a single point in ROC space, an ROC

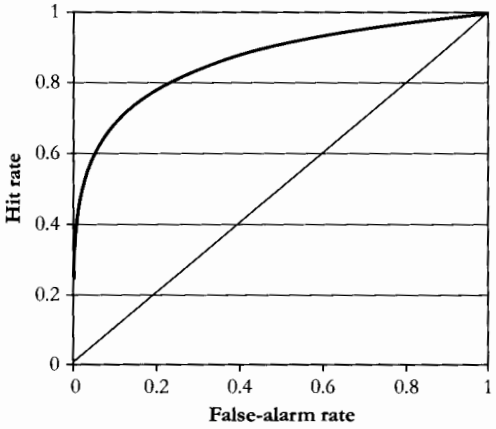

Figure 31.3 Receiver Operating Characteristic (ROC) describing signal detection performance.

cannot be estimated in an experiment in which only one such pair is contributed by a subject. Multiple pairs are usually estimated by manipulating response bias across within-subject conditions with a payoff or a base rate variable, as described previously, or by soliciting confidence ratings, which require the decision maker to maintain multiple response criteria simultaneously.

Multiple Decision Makers

Heretofore we have considered the behavior of a single decision maker, working in isolation. In many contexts, though, signal detection tasks are not performed by a single agent, but by collaborative teams of human and machine agents. A common example is an *alerted-monitor system*, in which an automated agent issues an alarm to notify a human decision maker when it detects that the status of a monitored subsystem has gone out of safe bounds (Sorkin & Woods, 1985). An electronic air traffic monitor, for instance, might detect and alert pilots to potential conflicts between aircraft. In such cases, performance of the joint human-machine system will be a function of four factors: the sensitivity and bias of the automated monitor, and the sensitivity and bias of the human decision maker. Ideally, performance of the joint system will exceed that of either of the individual component agents. If judgments made by the two agents are statistically independent and are weighted optimally to render a final decision, more specifically, the sensitivity of the joint system, as measured by the statistic d', will be $d'_{ha} = \sqrt{(d'^2_a + d'^2_h)}$, where d'_{ha} is the sensitivity of the human-automation system, d'_a is the sensitivity of the automated monitor by itself, and d'_h is the sensitivity of the unaided human operator (Pollack & Madans, 1964; Sorkin, Hays, & West, 2001). Unfortunately, empirical performance of automation-aided humans is rarely if ever optimal (Parasuraman & Riley, 1997) and often fails to even match the performance of the automation by itself (see Wickens & Dixon, 2007, for review). As this happens even when judgments of the automation and human are based on independent information sources, it indicates that human users typically fail to optimally combine their judgments with the automated aid's.

Chaining automated and human decision makers also complicates the problem of determining the automated agent's optimal criterion. As discussed above, sensitivity is independent of response criterion for a single decision maker. Within an alerted-monitor system, however, a change in the automated monitor's bias may not simply change the balance of the automation's hit and false-alarm rates, but may also change the human agent's willingness to trust the automation. Trust is a psychological process influenced by multiple factors (See & Lee, 2004), and the response criterion that optimizes an automated monitor's performance may not be the criterion that optimizes the human operator's trust or the behavior of the human-automation system as a whole (Sorkin & Woods, 1985). A variety of data show, for example, that holding all other factors equal, human operators tend to trust an aid that is biased toward misses more than they trust an aid biased toward false alarms (Dixon, Wickens, & McCarley, 2007). Thus, even if a bias toward false alarms is normative for the automated monitor considered by itself, it may not optimize sensitivity of the joint human-automation system. Such effects imply that until the behavior of the human operator can be predicted accurately from theoretical first principles, the process of determining the optimal criterion for the automated agent in a human-automation system will not be strictly analytical, but largely empirical.

Extensions of TSD

The mathematics of TSD are sufficiently straightforward that the theory has been generalized in ways that have offered greater explanatory power in understanding human behavior. The most prominent of these has been the development of multivariate TSD, which allows stimuli and responses to vary along multiple independent dimensions and thus brings a greater diversity of tasks into the domain of detection theoretic analysis. Multivariate TSD has been applied to prominent problems in perception (Ashby & Townsend, 1986) and memory (Banks, 2001); good overviews are provided by MacMillan and Creelman (2005) and Wickens (2002). More recent work in TSD has developed methods to isolate the effects of encoding noise from decision noise in detection performance (Benjamin et al., 2009), and has begun to formulate and apply a model of fuzzy signal detection (Masalonis & Parasuraman, 2003; Parasuraman, Masalonis, & Hancock, 2000) in which events and responses no longer fall into discrete categories, but can vary along a continuum from signal to noise.

Conclusion

Bayesian and signal detection models offer methods of making, modeling, and assessing judgment and decision making under uncertainty. While the

latter has a long history in disciplines related to cognitive engineering and the former has only recently come to fruition in these fields, both are under continued theoretic development. As noted above, ongoing work in TSD seeks to apply detection theoretic analysis to a broader range of problems than has been possible in the past, including problems for which the distinction between signal and noise is not clearly dichotomous, but fuzzy (Lu, Hinze, & Li, 2011; Parasuraman et al., 2003). Research using Bayesian network analysis, meanwhile, has begun to produce methods for real-time classification of human operator traits (Jipp et al., 2008) and states (Liang et al., 2007). Such methods stand to improve human safety and productivity in multiple ways, for example, by allowing an automated system to tailor its own behavior to the abilities and conditions of an operator (Byrne & Parasuraman, 1996) or by enabling an automated system to detect and alert an operator in a dangerous state of distraction (Lee, 2009). Developments like these will expand the scope and power of cognitive engineering, making Bayesian and signal detection models ever more valuable to the theory and practice of the discipline.

References

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93*, 154–179.

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology, 10*, 137–181.

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science, 11*, 267–273.

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language, 51*, 159–172.

Benjamin, A. S., Diaz, M. L., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116*(1), 84–115.

Byrne, E. A., & Parasuraman, R. (1996). Psychophysiology and adaptive automation. *Biological Psychology, 42*, 249–268.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*, 1–73.

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299*, 999–1001.

Craig, A. (1987). Signal detection theory and probability matching apply to vigilance. *Human Factors, 29*, 645–652.

Darwiche, A. (2009). *Modeling and reasoning with Bayesian networks.* Cambridge, England: Cambridge University Press.

Dinges, D. F., & Powell, J. W. (1985). Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments, & Computers, 17*, 652–655.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors, 49*, 564–572.

Engbert, R., & Krügel, A. (2010). Readers use Bayesian estimation for eye movement control. *Psychological Science, 21*, 366–371.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684–704.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* Oxford, England: John Wiley.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge handbook of computational cognitive modeling* (pp. 59–100). Cambridge, England: Cambridge University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17*, 767–773.

Gunzelmann, G., Gross, J. B., Gluck, K. A., & Dinges, D. F. (2009). Sleep deprivation and sustained attention performance: Integrating mathematical and cognitive modeling. *Cognitive Science*, 880–910.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 302–313.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine, 73*, 538–540.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science, 290*, 2261–2262.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 49*, 1295–1306.

Jipp, M., Badreddin, E., Abkai, C., & Hesser, J. (2008). Individual ability-based configuration: Cognitive profiling with Bayesian networks. *IEEE International Conference on Systems, Man and Cybernetics, SMC 2008*, 3359–3364.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430–454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models.* Cambridge, MA: MIT Press.

Lee, J. D. (2009). Can technology get your eyes back on the road? *Science, 324*, 344–346.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*, 50–80.

Liang, Y., Lee, J. D., & Reyes, M. L. (2007). Nonintrusive detection of driver cognitive distraction in real time using Bayesian networks. *Transportation Research Record, 2018*, 1–8.

Lu, Y., Hinze, J., & Li, Q. (2011). Developing fuzzy signal detection theory for workers' hazard perception measures on subway operations. *Safety Science, 49*, 491–497.

MacMillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide.* Mahwah, NJ: Erlbaum.

Masalonis, A. J., & Parasuraman, R. (2003). Fuzzy signal detection theory: Analysis of human and machine performance in air traffic control, and analytic considerations. *Ergonomics, 46*, 1045–1074.

Meissner, C. A., & Kassin, S. M. (2002). "He's guilty!": Investigator bias in judgments of truth and deception. *Law and Human Behavior, 26*, 469–480.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15*, 465–494

Parasuraman, R., Masalonis, A., & Hancock, P. (2000). Fuzzy signal detection theory: Basic postulates and formulas for analyzing human and machine performance. *Human Factors, 42*, 636–659.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*, 230–253.

Parks, T. E. (1966). Signal-detectability theory of recognition performance. *Psychological Review, 73*, 44–58.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A' and other misconceptions about signal detection theory. *Psychonomic Bulletin & Review, 10*, 556–569.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Mateo, CA: Morgan Kaufmann.

Pollack, I., & Madans, A. B. (1964). On the performance of a combination of detectors. *Human Factors, 6*, 523–531.

Rabinowitz, Y. G., Breitbach, J. E., & Warner, C. H. (2009). Managing aviator fatigue in a deployed environment: The relationship between fatigue and neurocognitive functioning. *Military Medicine, 174*, 358–362.

Renooij, S. (2001): Probability elicitation for belief networks: Issues to consider. *Knowledge Engineering Review, 16*, 255–269.

Rotello, C. M., & Macmillan, N. A. (2007). Response bias in recognition memory. In A. S. Benjamin & B. H. Ross (Eds.), *The psychology of learning and motivation: Skill and strategy in memory use* (Vol. 48, pp. 61–94). London, England: Academic Press.

Schoeffler, M. S. (1965). Theory for psychophysical learning. *Journal of the Acoustical Society of America, 37*, 1124–1133.

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review, 108*, 183–203.

Sivia, D., & Skilling, J. (2006). *Data analysis: A Bayesian tutorial* (2nd ed.). Oxford, England: Oxford University Press.

Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human-Computer Interaction, 1*, 49–75.

Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin, 99*, 181–198.

Treisman, M. (1987). Effects of the setting and adjustment of decision criteria on psychophysical performance. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 253–297). New York, NY: Elsevier Science.

Trucco, P., Cagno, E., Ruggeri, F., & Grande, O. (2008). A Bayesian belief network modelling of organisational factors in risk analysis: A case study of maritime transportation. *Reliability Engineering and System Safety, 93*, 823–834.

Vickers, D., & Leary, J. N. (1983). Criterion control in signal detection. *Human Factors, 25*, 283–296.

Wickelgren, W. A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. *Journal of Mathematical Psychology, 5*, 102–122.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science, 8*, 201–212.

Wickens, T. D. (2002). *Elementary signal detection theory.* New York, NY: Oxford University Press.