truths in the database without checking other truths that may reside there. The following discussion will focus on an even tougher problem, *query sensitivity*, which stems not from neglecting facts that were learned but from neglecting to specify which facts could have been learned. In other words, plausible reasoning, unlike logical deduction, is sensitive not only to the information at hand but also to the query process by which the information was obtained.

## *2.3.2 The Three Prisoners Paradox: When the Bare Facts Won't Do*

Three prisoners, *A*, *B*, and *C*, have been tried for murder, and their verdicts will be read and their sentences executed tomorrow morning. They know only that one of them will be declared guilty and will be hanged to die while the other two will be set free; the identity of the condemned prisoner is revealed to the very reliable prison guard, but not to the prisoners themselves.

In the middle of the night, Prisoner *A* calls the guard over and makes the following request: "Please give this letter to one of my friends—to one who is to be released. You and I know that at least one of them will be freed." The guard takes the letter and promises to do as told. An hour later Prisoner *A* calls the guard again and asks, "Can you tell me which of my friends you gave the letter to? It should give me no clue regarding my own status because, regardless of my fate, each of my friends had an equal chance of receiving my letter." The guard answers, "I gave the letter to Prisoner *B*; he will be released tomorrow." Prisoner *A* returns to his bed and thinks, "Before I talked to the guard, my chances of being executed were one in three. Now that he has told me that *B* will be released, only *C* and I remain, and my chances of dying have gone from 33.3% to 50%. What did I do wrong? I made certain not to ask for any information relevant to my own fate...."

### *SEARCHING FOR THE BARE FACTS*

So far, we have the classical Three Prisoners story as described in many books of mathematical puzzles (e.g., Gardner [1961]). Students are asked to test which of the two values, 1/3 or 1/2, reflects prisoner *A*'s updated chances of perishing at dawn.† Let us attempt to resolve the issue using formal probability theory.

† A survey conducted in the author's class in 1984 showed 23 students in favor of 1/2 and 3 students in favor of 1/3. (The proportion was reversed in 1987, when class notes became available.)

ıy reside there. The
n, *query sensitivity*,
t from neglecting to
plausible reasoning,
ın at hand but also to

## *hen*

their verdicts will be
now only that one of
the other two will be
to the very reliable

over and makes the
ds—to one who is to
e freed." The guard
oner *A* calls the guard
gave the letter to? It
egardless of my fate,
letter." The guard
morrow." Prisoner *A*
my chances of being
will be released, only
% to 50%. What did
relevant to my own

ed in many books of
sked to test which of
nces of perishing at
bility theory.

favor of 1/2 and 3 students
ecame available.)

Let $I_B$ stand for the proposition "Prisoner $B$ will be declared innocent," and let $G_A$ stand for the proposition "Prisoner $A$ will be declared guilty." Our task is to compute the probability of $G_A$ given all the information obtained from the guard, i.e., to compute $P(G_A \mid I_B)$. Since $G_A \supset I_B$, we have $P(I_B \mid G_A) = 1$, and we can write

$$P(G_A \mid I_B) = \frac{P(I_B \mid G_A)\, P(G_A)}{P(I_B)} = \frac{P(G_A)}{P(I_B)} = \frac{1/3}{2/3} = 1/2. \qquad \textbf{(2.55)}$$

Thus, when facts are wrongly formulated, even the tools of probability calculus are insufficient safeguards against drawing counterintuitive or false conclusions. (Readers who are not convinced that the answer 50% is false are invited to eavesdrop on Prisoner $A$'s further reflections: "... Worse yet, by sheer symmetry, my chances of dying would also have risen to 50% if the guard had named $C$ instead of $B$—so my chances must have been 50% to begin with. I must be hallucinating....")

The fallacy in the preceding formulation arose from omitting the full context in which the answer was obtained by Prisoner $A$. By *context* we mean the entire range of answers one could possibly obtain (as in Eq. (2.30)), not just the answer actually obtained. In our example, it is important to know not only that the guard said, "$B$ will be released," but also that the only other possible reply was "$C$ will be released." Had the guard's answer, "$B$ will be released," been a reply to the query "Will $B$ die tomorrow?" the preceding analysis would have been correct.

A useful way of ensuring that we have considered the full context is to condition our analysis on events actually observed, not on their implications. In our example, the information in

$$I_B = \text{"}B \text{ will be declared innocent."}$$

was inferred from a more direct observation,

$$I'_B = \text{"Guard said that } B \text{ will be declared innocent."}$$

If we compute $P(G_A \mid I'_B)$ instead of $P(G_A \mid I_B)$, we get the correct answer:

$$P(G_A \mid I'_B) = \frac{P(I'_B \mid G_A) P(G_A)}{P(I'_B)} = \frac{1/2 \cdot 1/3}{1/2} = 1/3. \qquad \textbf{(2.56)}$$

The calculations in Eq. (2.56) differ from those in Eq. (2.55) in two ways. First, $G_A$ subsumed $I_B$ but does not subsume $I'_B$, because it is possible for $A$ to be the condemned man and hear the guard report, "$C$ will be released." Second, $P(I'_B)$ is 1/2, whereas $P(I_B)$ was 2/3. These differences exist because $I'_B$ implies $I_B$ but not vice versa; even if $B$ is to be released, the guard can truthfully report, "$C$ will be released"—if $A$ is slated to die.

The lesson of the Three Prisoners paradox is that we cannot assess the impact of new information by considering only propositions implied by the information; we must also consider what information *could have* been reported.

## *THE THOUSAND PRISONER PROBLEM*

Here is an extreme example, in which knowledge of the query context is even more important. Imagine you are one of one thousand prisoners awaiting sentencing with the knowledge that only one of you has been condemned. By sheer luck, you find a computer printout (with a court seal on it) listing 998 prisoners; each name is marked "innocent," and yours is not among them. Should your chances of dying increase from 1/1000 to 1/2? Most people would say yes, and rightly so.

Imagine, however, that while poring anxiously over the list you discover the query that produced it: "Print the names of any 998 innocent right-handed prisoners." If you are the only left-handed person around, would you not breathe a sigh of relief? Again, most people would.

Though the discovery of the query adds no logical conclusions to our knowledge base, it alters drastically the relative likelihood of events that remain unsettled. In other words, the range of possibilities is the same before and after you discover the query: Either you or the other unlisted prisoner will die. Yet the query renders the death of the other prisoner much more likely, because while you can blame your exclusion from the list on being left-handed, the other prisoner has no explanation except being found guilty. If the list contained 999 names marked "innocent," knowledge of the query would have no impact on your beliefs, because the only possible conclusion would be that you had been found guilty.

Again we see the computational virtues and epistemological weaknesses of crisp logic: It allows us to dispose of the query once we learn its ramifications but prevents the ramifications learned from altering the likelihood of uncertain events. Indeed, if we wish to determine merely which events are possible we need not retain the queries; the bare information will suffice. But if we are concerned also with the relative likelihood of these possible events, then the query process is necessary. If the process is unknown, then several likely processes can be conjectured and their average computed (see next subsection).

But first, let us return to the jail cell. Mathematically, the discovery of the query should restore your confidence of innocence to its original value of 99.9%, but psychologically you are more frightened than you were before you found the list. In your intuition, the realization that you are one of the only two potentially guilty individuals evidently carries more weight than Bayesian arithmetic does. Still, intuition is a multifaceted resource, and pondering further, you should muster intuitive support for the Bayesian conclusion as well: Finding the query after seeing the list should have the same effect as seeing the list after the query. In the second case, once you know the query, the list is useless to you, because it can

ssess the impact
the information;

context is even
isoners awaiting
condemned. By
n it) listing 998
ng them. Should
le would say yes,

you discover the
ent right-handed
you not breathe a

nclusions to our
vents that remain
before and after
will die. Yet the
ecause while you
other prisoner has
99 names marked
r beliefs, because
ilty.
al weaknesses of
ramifications but
uncertain events.
sible we need not
re concerned also
query process is
processes can be

discovery of the
l value of 99.9%,
ore you found the
ly two potentially
n arithmetic does.
you should muster
g the query after
r the query. In the
ou, because it can

contain neither your name nor the name of the guilty prisoner. Consequently, your chances of being found guilty should revert to 1/1000.

## WHAT IF WE DON'T KNOW THE QUERY?

In the Three Prisoners story, we assumed that if both $B$ and $C$ were pardoned, the guard would give the letter to one or the other with equal (½) probability. What if we do not know the process by which the letter recipient is chosen, when $A$ is condemned? The conditional probability $P(I'_B|G_A)$ can vary from 0 (the guard avoids $B$), to 1 (the guard avoids $C$). Likewise, the marginal probability $P(I'_B)$ can vary from $^1/_3$ to $^2/_3$. Treating $q = P(I'_B|G_A)$ as a variable, Eq. (2.56) can be written as follows:

$$P(G_A|I'_B) = \frac{P(I'_B|G_A)\ P(G_A)}{P(I'_B|G_A)\ P(G_A) + P(I'_B|G_B)\ P(G_B) + P(I'_B|G_C)\ P(G_C)}$$

$$= \frac{q\ ^1/_3}{q\ ^1/_3 + 0 + 1 \cdot {}^1/_3} = \frac{q}{1+q}\ . \tag{2.57}$$

Thus, as $q$ varies from 0 to 1, $P(G_A|I'_B)$ varies from 0 to ½.

Philosophers disagree on how to treat ignorance of this sort. Some favor the use of probability intervals, where the upper and lower probabilities represent the boundaries of our convictions, while others prefer an interpolation rule that selects a single probability model having some desirable properties. The Dempster-Shafer (D-S) formalism (see Chapter 9) is an example of the interval-based approach, while maximum-entropy techniques [Tribus 1969, Jaynes 1979] represent the single model approach.

Bayesian technique lies somewhere in between. For example, in the absence of information about the selection process used by the guard, several plausible models of the process are articulated, and their likelihoods are assessed. In our example, we may treat the critical parameter $q$ as a random variable ranging from 0 to 1 and assess a probability distribution $f(q)$ on $q$, reflecting the likelihood that the guard will exhibit a bias $q$ in favor of selecting $B$. This method yields a unique distribution on the variables previously considered, via

$$P(G_A|I'_B) = \int_0^1 \frac{q}{1+q}\ f(q|I'_B)\ dq = \frac{\int_0^1 q\ f\ (q)dq}{1 + \int_0^1 q\ f\ (q)dq}\ , \tag{2.58}$$

but the method simultaneously maintains a distinction between conclusions based on definite models and conclusions based on uncertain models. For example, the knowledge that the choice between $B$ and $C$ is made at random is modeled by $q = ½$, while total lack of knowledge about the process is represented by $f(q) = 1\ ,\ 0 \le q \le 1$. Though both models yield the same point values of $^1/_3$ for