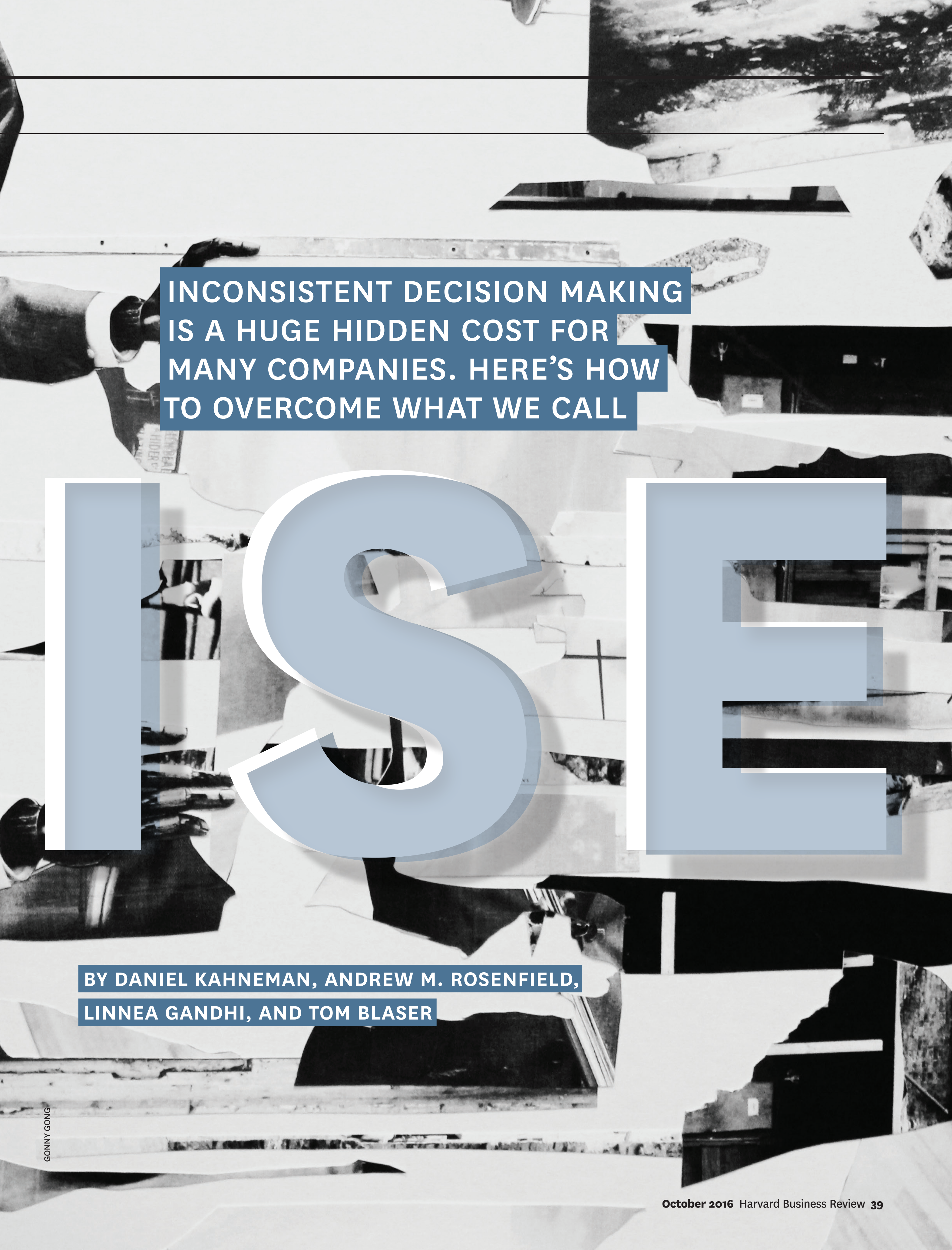
THE BIG IDEA

Daniel Kahneman is the Eugene Higgins Professor of Psychology Emeritus at Princeton University. He was awarded the Nobel Prize in Economic Sciences in 2002 for his work (with Amos Tversky) on cognitive biases. **Andrew M. Rosenfield** is the CEO and managing partner of The Greatest Good Group (TGG Group). **Linnea Gandhi** and **Tom Blaser** are managing directors at TGG Group.



INCONSISTENT DECISION MAKING IS A HUGE HIDDEN COST FOR MANY COMPANIES. HERE'S HOW TO OVERCOME WHAT WE CALL

ISE

BY DANIEL KAHNEMAN, ANDREW M. ROSENFIELD, LINNEA GANDHI, AND TOM BLASER

GONNY GONG

At a global financial services firm we worked with, a longtime customer accidentally submitted the same application file to two offices. Though the employees who reviewed the file were supposed to follow the same guidelines—and thus arrive at similar outcomes—the separate offices returned very different quotes. Taken aback, the customer gave the business to a competitor. From the point of view of the firm, employees in the same role should have been interchangeable, but in this case they were not. Unfortunately, this is a common problem.

Professionals in many organizations are assigned arbitrarily to cases: appraisers in credit-rating agencies, physicians in emergency rooms, underwriters of loans and insurance, and others. Organizations expect consistency from these professionals: Identical cases should be treated similarly, if not identically. The problem is that humans are unreliable decision makers; their judgments are strongly influenced by irrelevant factors, such as their current mood, the time since their last meal, and the weather. We call the chance variability of judgments *noise*. It is an invisible tax on the bottom line of many companies.

Some jobs are noise-free. Clerks at a bank or a post office perform complex tasks, but they must follow strict rules that limit subjective judgment and guarantee, by design, that identical cases will be treated identically. In contrast, medical professionals, loan officers, project managers, judges, and executives all make judgment calls, which are guided by informal experience and general principles rather than by rigid rules. And if they don't reach precisely the same answer that every other person in their role would, that's acceptable; this is what we mean when we say that a decision is "a matter of judgment." A firm whose employees exercise judgment does not expect decisions to be entirely free of noise. But often noise is *far above* the level that executives would consider tolerable—and they are completely unaware of it.

The prevalence of noise has been demonstrated in several studies. Academic researchers have repeatedly confirmed that professionals often contradict their own prior judgments when given the same data on different occasions. For instance, when software developers were asked on two separate days to estimate the completion time for a given task, the hours they projected differed by 71%, on average. When pathologists made two assessments of the severity of biopsy results, the correlation between their ratings was only .61 (out of a perfect 1.0), indicating that they made inconsistent diagnoses quite frequently. Judgments made by different people are even more likely to diverge. Research has confirmed that in many tasks, experts' decisions are highly variable: valuing stocks, appraising real estate, sentencing criminals, evaluating job performance, auditing financial statements, and more. The unavoidable conclusion is that professionals often make decisions that deviate significantly from those of their peers, from their own prior decisions, and from rules that they themselves claim to follow.

Noise is often insidious: It causes even successful companies to lose substantial amounts of money without realizing it. How substantial? To get an estimate, we asked executives in one of the organizations we studied the following: "Suppose the optimal assessment of a case is $100,000. What would be the cost to the organization if the professional in charge of the case assessed a value of $115,000? What would be the cost of assessing it at $85,000?" The cost estimates were high. Aggregated over the assessments made every year, the cost of noise was measured in billions—an unacceptable number even for a large global firm. The value of reducing noise even by a few percentage points would be in

Idea in Brief

THE PROBLEM

Many organizations expect consistency from their professional employees. However, human judgment is often influenced by such irrelevant factors as the weather and the last case seen. More important, decisions often vary from employee to employee. The chance variability of judgments is called *noise,* and it is surprisingly costly to companies.

THE STARTING POINT

Managers should perform a noise audit in which members of a unit, working independently, evaluate a common set of cases. The degree to which their decisions vary is the measure of noise. It will often be dramatically higher than executives anticipate.

THE SOLUTION

The most radical solution to a severe noise problem is to replace human judgment with algorithms. Algorithms are not difficult to construct—but often they're politically or operationally infeasible. In such instances, companies should establish procedures to help professionals achieve greater consistency.

---

the tens of millions. Remarkably, the organization had completely ignored the question of consistency until then.

It has long been known that predictions and decisions generated by simple statistical algorithms are often more accurate than those made by experts, even when the experts have access to more information than the formulas use. It is less well known that the key advantage of algorithms is that they are noise-free: Unlike humans, a formula will always return the same output for any given input. Superior consistency allows even simple and imperfect algorithms to achieve greater accuracy than human professionals. (Of course, there are times when algorithms will be operationally or politically infeasible, as we will discuss.)

In this article we explain the difference between noise and bias and look at how executives can audit the level and impact of noise in their organizations. We then describe an inexpensive, underused method for building algorithms that remediate noise, and we sketch out procedures that can promote consistency when algorithms are not an option.

Noise vs. Bias

When people consider errors in judgment and decision making, they most likely think of social biases like the stereotyping of minorities or of cognitive biases such as overconfidence and unfounded optimism. The useless variability that we call noise is a different type of error. To appreciate the distinction, think of your bathroom scale. We would say that the scale is *biased* if its readings are generally either too high or too low. If your weight appears to depend on where you happen to place your feet, the scale is *noisy.* A scale that consistently underestimates true weight by exactly four pounds is seriously biased but free of noise. A scale that gives two different readings when you step on it twice is noisy. Many errors of measurement arise from a combination of bias and noise. Most inexpensive bathroom scales are somewhat biased and quite noisy.

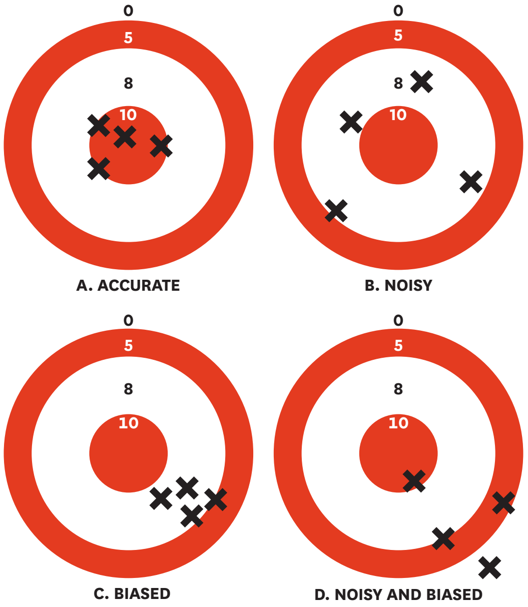
For a visual illustration of the distinction, consider the targets in the exhibit "How Noise and Bias Affect Accuracy." These show the results of target practice for four-person teams in which each individual shoots once.

- Team A is *accurate:* The shots of the teammates are on the bull's-eye and close to one another.

The other three teams are inaccurate but in distinctive ways:

- Team B is *noisy:* The shots of its members are centered around the bull's-eye but widely scattered.

HOW NOISE AND BIAS AFFECT ACCURACY



A. ACCURATE

B. NOISY

C. BIASED

D. NOISY AND BIASED

- Team C is *biased:* The shots all missed the bull's-eye but cluster together.
- Team D is both *noisy* and *biased*.

As a comparison of teams A and B illustrates, an increase in noise always impairs accuracy when there is no bias. When bias is present, increasing noise may actually cause a lucky hit, as happened for team D. Of course, no organization would put its trust in luck. Noise is always undesirable—and sometimes disastrous.

It is obviously useful to an organization to know about bias and noise in the decisions of its employees, but collecting that information isn't straightforward. Different issues arise in measuring these errors. A major problem is that the outcomes of decisions often aren't known until far in the future, if at all. Loan officers, for example, frequently must wait several years to see how loans they approved worked out, and they almost never know what happens to an applicant they reject.

Unlike bias, noise can be measured without knowing what an accurate response would be. To illustrate, imagine that the targets at which the shooters aimed were erased from the exhibit. You would know nothing about the teams' overall accuracy, but you could be certain that something was wrong with the scattered shots of teams B and D: Wherever the bull's-eye was, they did not all come close to hitting it. All that's required to measure noise in judgments is a simple experiment in which a few realistic cases are evaluated independently by several professionals. Here again, the scattering of judgments can be observed without knowing the correct answer. We call such experiments *noise audits.*

WHERE THERE IS JUDGMENT, THERE IS NOISE—AND USUALLY MORE OF IT THAN YOU THINK.

Performing a Noise Audit

The point of a noise audit is not to produce a report. The ultimate goal is to improve the quality of decisions, and an audit can be successful only if the leaders of the unit are prepared to accept unpleasant results and act on them. Such buy-in is easier to achieve if the executives view the study as their own creation. To that end, the cases should be compiled by respected team members and should cover the range of problems typically encountered. To make the results relevant to everyone, all unit members should participate in the audit. A social scientist with experience in conducting rigorous behavioral experiments should supervise the technical aspects of the audit, but the professional unit must own the process.

Recently, we helped two financial services organizations conduct noise audits. The duties and expertise of the two groups we studied were quite different, but both required the evaluation of moderately complex materials and often involved decisions about hundreds of thousands of dollars. We followed the same protocol in both organizations. First we asked managers of the professional teams involved to construct several realistic case files for evaluation. To prevent information about the experiment from leaking, the entire exercise was conducted on the same day. Employees were asked to spend about half the day analyzing two to four cases. They were to decide on a dollar amount for each, as in their normal routine. To avoid collusion, the participants were not told that the study was concerned with reliability. In one organization, for example, the goals were described as understanding the employees' professional thinking, increasing their tools' usefulness, and improving communication among colleagues. About 70 professionals in organization A participated, and about 50 in organization B.

We constructed a noise index for each case, which answered the following question: "By how much do the judgments of two randomly chosen employees differ?" We expressed this amount as a percentage of their average. Suppose the assessments of a case by two employees are $600 and $1,000. The average of their assessments is $800, and the difference between them is $400, so the noise index is 50% for this pair. We performed the same computation for all pairs of employees and then calculated an overall average noise index for each case.

Pre-audit interviews with executives in the two organizations indicated that they expected the differences between their professionals' decisions to range from 5% to 10%—a level they considered acceptable for "matters of judgment." The results came as a shock. The noise index ranged from 34% to 62% for the six cases in organization A, and the overall average was 48%. In the four cases in organization B, the noise index ranged from 46% to 70%, with an average of 60%. Perhaps most disappointing, experience on the job did not appear to reduce noise. Among professionals with five or more years on the job, average disagreement was 46% in organization A and 62% in organization B.

No one had seen this coming. But because they owned the study, the executives in both organizations accepted the conclusion that the judgments of their professionals were unreliable to an extent that could not be tolerated. All quickly agreed that something had to be done to control the problem.

Because the findings were consistent with prior research on the low reliability of professional judgment, they didn't surprise us. The major puzzle for us was the fact that neither organization had ever considered reliability to be an issue.

The problem of noise is effectively invisible in the business world; we have observed that audiences are quite surprised when the reliability of professional judgment is mentioned as an issue. What prevents companies from recognizing that the judgments of their employees are noisy? The answer lies in two familiar phenomena: Experienced professionals tend to have high confidence in the accuracy of their own judgments, and they also have high regard for their colleagues' intelligence. This combination inevitably leads to an overestimation of agreement. When asked about what their colleagues would say, professionals expect others' judgments to be much closer to their own than they actually are. Most of the time, of course, experienced professionals are completely unconcerned with what others might think and simply assume that theirs is the best answer. One reason the problem of noise is invisible is that people do not go through life imagining plausible alternatives to every judgment they make.

The expectation that others will agree with you is sometimes justified, particularly where judgments are so skilled that they are intuitive. High-level chess and driving are standard examples of tasks that have been practiced to near perfection. Master players who look at a situation on a chessboard will all have very similar assessments of the state of the game—whether, say, the white queen is in danger or black's king-side defense is weak. The same is true of drivers. Negotiating traffic would be impossibly dangerous if we could not assume that the drivers around us share our understanding of priorities at intersections and roundabouts. There is little or no noise at high levels of skill.

Types of Noise and Bias

Bias and noise are distinct kinds of error. Each comes in different variants and requires different corrective actions.

TYPE OF BIAS	EXAMPLES	CORRECTIVE ACTIONS
GENERAL The average judgment is wrong.	• Planning fallacy: Forecasts of outcomes are mostly optimistic • Excessive risk aversion: A venture capital firm rejects too many promising but risky investments	• Continual monitoring of decisions • Guidelines and targets for the frequency of certain outcomes (such as loan approvals) • Eliminating incentives that favor biases
SOCIAL Discrimination occurs against—or for—certain categories of cases.	• Frequent denial of credit to qualified applicants from certain ethnic groups • Gender bias in assessments of job performance	• Monitoring statistics for different groups • Blinding of applications • Objective and quantifiable metrics • Open channels for complaints • Guidelines and training
COGNITIVE Decisions are strongly influenced by irrelevant factors or insensitive to relevant ones.	• Excessive effects of first impressions • Effects of anchors (such as an opening offer in negotiation) • Myopic neglect of future consequences	• Training employees to detect situations in which biases are likely to occur • Critiques of important decisions, focused on likely biases

TYPE OF NOISE	EXAMPLES	CORRECTIVE ACTIONS
VARIABILITY ACROSS OCCASIONS Decisions vary when the same case is presented more than once to the same individual.	• A hiring officer's judgments of a file are influenced by her mood or the quality of the previous applicant	• Algorithms to replace human judgment • Checklists that encourage a consistent approach to decisions
VARIABILITY ACROSS INDIVIDUALS Professionals in the same role make different decisions.	• Some individuals are generally more lenient than others • Some individuals are more cautious than others	• Algorithms to replace human judgment • Frequent monitoring of individuals' decisions • Roundtables at which differences are explored and resolved • Checklists that encourage a consistent approach to decisions

High skill develops in chess and driving through years of practice in a predictable environment, in which actions are followed by feedback that is both immediate and clear. Unfortunately, few professionals operate in such a world. In most jobs people learn to make judgments by hearing managers and colleagues explain and criticize—a much less reliable source of knowledge than learning from one's mistakes. Long experience on a job always increases people's confidence in their judgments, but in the absence of rapid feedback, confidence is no guarantee of either accuracy or consensus.

We offer this aphorism in summary: *Where there is judgment, there is noise—and usually more of it than you think.* As a rule, we believe that neither professionals nor their managers can make a good guess about the reliability of their judgments. The only way to get an accurate assessment is to conduct a noise audit. And at least in some cases, the problem will be severe enough to require action.

Dialing Down the Noise

The most radical solution to the noise problem is to replace human judgment with formal rules—known as algorithms—that use the data about a case to produce a prediction or a decision. People have competed against algorithms in several hundred contests of accuracy over the past 60 years, in tasks ranging from predicting the life expectancy of cancer patients to predicting the success of graduate students. Algorithms were more accurate than human professionals in about half the studies, and approximately tied with the humans in the others. The ties should also count as victories for the algorithms, which are more cost-effective.

In many situations, of course, algorithms will not be practical. The application of a rule may not be feasible when inputs are idiosyncratic or hard to code in a consistent format. Algorithms are also less likely to be useful for judgments or decisions that involve multiple dimensions or depend on negotiation with another party. Even when an algorithmic solution is available in principle, organizational considerations sometimes prevent implementation. The replacement of existing employees by software is a painful process that will encounter resistance unless it frees those employees up for more-enjoyable tasks.

But if the conditions are right, developing and implementing algorithms can be surprisingly easy. The common assumption is that algorithms require statistical analysis of large amounts of data. For example, most people we talk to believe that data on thousands of loan applications and their outcomes is needed to develop an equation that predicts commercial loan defaults. Very few know that adequate algorithms can be developed without any outcome data at all—and with input information on only a small number of cases. We call predictive formulas that are built without outcome data "reasoned rules," because they draw on commonsense reasoning.

STUDIES HAVE SHOWN THAT WHILE HUMANS CAN PROVIDE USEFUL INPUT, ALGORITHMS DO BETTER IN THE ROLE OF FINAL DECISION MAKER.

The construction of a reasoned rule starts with the selection of a few (perhaps six to eight) variables that are incontrovertibly related to the outcome being predicted. If the outcome is loan default, for example, assets and liabilities will surely be included in the list. The next step is to assign these variables equal weight in the prediction formula, setting their sign in the obvious direction (positive for assets, negative for liabilities). The rule can then be constructed by a few simple calculations. (For more details, see the sidebar "How to Build a Reasoned Rule.")

The surprising result of much research is that in many contexts reasoned rules are about as accurate as statistical models built with outcome data. Standard statistical models combine a set of predictive variables, which are assigned weights based on their relationship to the predicted outcomes and to one another. In many situations, however, these weights are both statistically unstable and practically unimportant. A simple rule that assigns equal weights to the selected variables is likely to be just as valid. Algorithms that weight variables equally and don't rely on outcome data have proved successful in personnel selection, election forecasting, predictions about football games, and other applications.

The bottom line here is that if you plan to use an algorithm to reduce noise, you need not wait for outcome data. You can reap most of the benefits by using common sense to select variables and the simplest possible rule to combine them.

Of course, no matter what type of algorithm is employed, people must retain ultimate control. Algorithms must be monitored and adjusted for occasional changes in the population of cases. Managers must also keep an eye on individual decisions and have the authority to override the algorithm in clear-cut cases. For example, a decision to approve a loan should be provisionally reversed if the firm discovers that the applicant has been arrested. Most important, executives should determine how to translate the algorithm's output into action. The algorithm can tell you which prospective loans are in the top 5% or in the bottom 10% of all applications, but someone must decide what to do with that information.

Algorithms are sometimes used as an intermediate source of information for professionals, who make the final decisions. One example is the Public Safety Assessment, a formula that was developed to help U.S. judges decide whether a defendant can be safely released pending trial. In its first six months of use in Kentucky, crime among defendants on pretrial release fell by about 15%, while the percentage of people released pretrial increased. It's obvious in this case that human judges must retain the final authority for the decisions: The public would be shocked to see justice meted out by a formula.

Uncomfortable as people may be with the idea, studies have shown that while humans can provide useful input to formulas, algorithms do better in the role of final decision maker. If the avoidance of errors is the only criterion, managers should be strongly advised to overrule the algorithm only in exceptional circumstances.

Bringing Discipline to Judgment

Replacing human decisions with an algorithm should be considered whenever professional judgments are noisy, but in most cases this solution will be too radical or simply impractical. An alternative is to adopt procedures that promote consistency by ensuring that employees in the same role use similar methods to seek information, integrate it into a view of the case, and translate that view into a decision. A thorough examination of everything required to do that is beyond the scope of this article, but we can offer some basic advice, with the important caveat that instilling discipline in judgment is not at all easy.

Training is crucial, of course, but even professionals who were trained together tend to drift into their own way of doing things. Firms sometimes combat drift by organizing roundtables at which decision makers gather to review cases. Unfortunately, most roundtables are run in a way that makes it much too easy to achieve agreement, because participants quickly converge on the opinions stated first or most confidently. To prevent such spurious agreement, the individual participants in a roundtable should study the case independently, form opinions they're prepared to defend, and send those opinions to the group leader before the meeting. Such roundtables will effectively provide an audit of noise, with the added step of a group discussion in which differences of opinion are explored.

As an alternative or addition to roundtables, professionals should be offered user-friendly tools, such as checklists and carefully formulated questions, to guide them as they collect information about a case, make intermediate judgments, and formulate a final decision. Unwanted variability occurs at each of those stages, and firms can—and should—test how much such tools reduce it. Ideally, the people who use these tools will view them as aids that help them do their jobs effectively and economically.

How to Build a Reasoned Rule

You don't need outcome data to create useful predictive algorithms. For example, you can build a reasoned rule that predicts loan defaults quite effectively without knowing what happened to past loans; all you need is a small set of recent loan applications. Here are the next steps:

1 Select six to eight variables that are distinct and obviously related to the predicted outcome. Assets and revenues (weighted positively) and liabilities (weighted negatively) would surely be included, along with a few other features of loan applications.

2 Take the data from your set of cases (all the loan applications from the past year) and compute the mean and standard deviation of each variable in that set.

3 For every case in the set, compute a "standard score" for each variable: the difference between the value in the case and the mean of the whole set, divided by the standard deviation. With standard scores, all variables are expressed on the same scale and can be compared and averaged.

4 Compute a "summary score" for each case—the average of its variables' standard scores. This is the output of the reasoned rule. The same formula will be used for new cases, using the mean and standard deviation of the original set and updating periodically.

5 Order the cases in the set from high to low summary scores, and determine the appropriate actions for different ranges of scores. With loan applications, for instance, the actions might be "the top 10% of applicants will receive a discount" and "the bottom 30% will be turned down."

You are now ready to apply the rule to new cases. The algorithm will compute a summary score for each new case and generate a decision.

Unfortunately, our experience suggests that the task of constructing judgment tools that are both effective and user-friendly is more difficult than many executives think. Controlling noise is hard, but we expect that an organization that conducts an audit and evaluates the cost of noise in dollars will conclude that reducing random variability is worth the effort.

OUR MAIN GOAL in this article is to introduce managers to the concept of noise as a source of errors and explain how it is distinct from bias. The term "bias" has entered the public consciousness to the extent that the words "error" and "bias" are often used interchangeably. In fact, better decisions are not achieved merely by reducing general biases (such as optimism) or specific social and cognitive biases (such as discrimination against women or anchoring effects). Executives who are concerned with accuracy should also confront the prevalence of inconsistency in professional judgments. Noise is more difficult to appreciate than bias, but it is no less real or less costly.

HBR Reprint R1610B

Copyright © Harvard Business Publishing. All Rights Reserved. This content is intended for individual research use only, subject to the following:

Unless permission is expressly granted in a separate license, this content may NOT be used for classroom or teaching use, which includes teaching materials, electronic reserves, course packs or persistent linking from syllabi. Please consult your institution's librarian about the nature of relevant licenses held by your institution and the restrictions that may or may not apply.

Unless permission is expressly granted in a separate license, this content may NOT be used in corporate training and/or as corporate learning materials. For corporate users, please consult the specific terms of your company's license(s) for complete information and restrictions.

For more information and teaching resources from Harvard Business Publishing including Harvard Business School Cases, eLearning products, and business simulations please visit hbsp.harvard.edu.