21

INTUITIONS VS. FORMULAS

Paul Meehl was a strange and wonderful character, and one of the most versatile psychologists of the twentieth century. Among the departments in which he had faculty appointments at the University of Minnesota were psychology, law, psychiatry, neurology, and philosophy. He also wrote on religion, political science, and learning in rats. A statistically sophisticated researcher and a fierce critic of empty claims in clinical psychology, Meehl was also a practicing psychoanalyst. He wrote thoughtful essays on the philosophical foundations of psychological research that I almost memorized while I was a graduate student. I never met Meehl, but he was one of my heroes from the time I read his *Clinical vs. Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.*

In the slim volume that he later called "my disturbing little book," Meehl reviewed the results of 20 studies that had analyzed whether *clinical predictions* based on the subjective impressions of trained professionals were more accurate than *statistical* predictions made by combining a few scores or ratings according to a rule. In a typical study, trained counselors predicted the grades of freshmen at the end of the school year. The counselors interviewed each student for forty-five minutes. They also had access to high school grades, several aptitude tests, and a four-page personal statement. The statistical algorithm used only a fraction of this information: high school grades and one aptitude test. Nevertheless, the formula was more accurate than 11 of the 14 counselors. Meehl reported generally sim-

ilar results across a variety of other forecast outcomes, including violations of parole, success in pilot training, and criminal recidivism.

Not surprisingly, Meehl's book provoked shock and disbelief among clinical psychologists, and the controversy it started has engendered a stream of research that is still flowing today, more than fifty years after its publication. The number of studies reporting comparisons of clinical and statistical predictions has increased to roughly two hundred, but the score in the contest between algorithms and humans has not changed. About 60% of the studies have shown significantly better accuracy for the algorithms. The other comparisons scored a draw in accuracy, but a tie is tantamount to a win for the statistical rules, which are normally much less expensive to use than expert judgment. No exception has been convincingly documented.

The range of predicted outcomes has expanded to cover medical variables such as the longevity of cancer patients, the length of hospital stays, the diagnosis of cardiac disease, and the susceptibility of babies to sudden infant death syndrome; economic measures such as the prospects of success for new businesses, the evaluation of credit risks by banks, and the future career satisfaction of workers; questions of interest to government agencies, including assessments of the suitability of foster parents, the odds of recidivism among juvenile offenders, and the likelihood of other forms of violent behavior; and miscellaneous outcomes such as the evaluation of scientific presentations, the winners of football games, and the future prices of Bordeaux wine. Each of these domains entails a significant degree of uncertainty and unpredictability. We describe them as "low-validity environments." In every case, the accuracy of experts was matched or exceeded by a simple algorithm.

As Meehl pointed out with justified pride thirty years after the publication of his book, "There is no controversy in social science which shows such a large body of qualitatively diverse studies coming out so uniformly in the same direction as this one."

The Princeton economist and wine lover Orley Ashenfelter has offered a compelling demonstration of the power of simple statistics to outdo world-renowned experts. Ashenfelter wanted to predict the future value of fine Bordeaux wines from information available in the year they are made. The question is important because fine wines take years to reach their peak quality, and the prices of mature wines from the same vineyard vary dramatically across different vintages; bottles filled only twelve months apart

can differ in value by a factor of 10 or more. An ability to forecast future prices is of substantial value, because investors buy wine, like art, in the anticipation that its value will appreciate.

It is generally agreed that the effect of vintage can be due only to variations in the weather during the grape-growing season. The best wines are produced when the summer is warm and dry, which makes the Bordeaux wine industry a likely beneficiary of global warming. The industry is also helped by wet springs, which increase quantity without much effect on quality. Ashenfelter converted that conventional knowledge into a statistical formula that predicts the price of a wine—for a particular property and at a particular age—by three features of the weather: the average temperature over the summer growing season, the amount of rain at harvesttime, and the total rainfall during the previous winter. His formula provides accurate price forecasts years and even decades into the future. Indeed, his formula forecasts future prices much more accurately than the current prices of young wines do. This new example of a "Meehl pattern" challenges the abilities of the experts whose opinions help shape the early price. It also challenges economic theory, according to which prices should reflect all the available information, including the weather. Ashenfelter's formula is extremely accurate—the correlation between his predictions and actual prices is above .90.

Why are experts inferior to algorithms? One reason, which Meehl suspected, is that experts try to be clever, think outside the box, and consider complex combinations of features in making their predictions. Complexity may work in the odd case, but more often than not it reduces validity. Simple combinations of features are better. Several studies have shown that human decision makers are inferior to a prediction formula even when they are given the score suggested by the formula! They feel that they can overrule the formula because they have additional information about the case, but they are wrong more often than not. According to Meehl, there are few circumstances under which it is a good idea to substitute judgment for a formula. In a famous thought experiment, he described a formula that predicts whether a particular person will go to the movies tonight and noted that it is proper to disregard the formula if information is received that the individual broke a leg today. The name "broken-leg rule" has stuck. The point, of course, is that broken legs are very rare—as well as decisive.

Another reason for the inferiority of expert judgment is that humans are incorrigibly inconsistent in making summary judgments of complex information. When asked to evaluate the same information twice, they fre-

quently give different answers. The extent of the inconsistency is often a matter of real concern. Experienced radiologists who evaluate chest X-rays as "normal" or "abnormal" contradict themselves 20% of the time when they see the same picture on separate occasions. A study of 101 independent auditors who were asked to evaluate the reliability of internal corporate audits revealed a similar degree of inconsistency. A review of 41 separate studies of the reliability of judgments made by auditors, pathologists, psychologists, organizational managers, and other professionals suggests that this level of inconsistency is typical, even when a case is reevaluated within a few minutes. Unreliable judgments cannot be valid predictors of anything.

The widespread inconsistency is probably due to the extreme context dependency of System 1. We know from studies of priming that unnoticed stimuli in our environment have a substantial influence on our thoughts and actions. These influences fluctuate from moment to moment. The brief pleasure of a cool breeze on a hot day may make you slightly more positive and optimistic about whatever you are evaluating at the time. The prospects of a convict being granted parole may change significantly during the time that elapses between successive food breaks in the parole judges' schedule. Because you have little direct knowledge of what goes on in your mind, you will never know that you might have made a different judgment or reached a different decision under very slightly different circumstances. Formulas do not suffer from such problems. Given the same input, they always return the same answer. When predictability is poor—which it is in most of the studies reviewed by Meehl and his followers—inconsistency is destructive of any predictive validity.

The research suggests a surprising conclusion: to maximize predictive accuracy, final decisions should be left to formulas, especially in low-validity environments. In admission decisions for medical schools, for example, the final determination is often made by the faculty members who interview the candidate. The evidence is fragmentary, but there are solid grounds for a conjecture: conducting an interview is likely to diminish the accuracy of a selection procedure, if the interviewers also make the final admission decisions. Because interviewers are overconfident in their intuitions, they will assign too much weight to their personal impressions and too little weight to other sources of information, lowering validity. Similarly, the experts who evaluate the quality of immature wine to predict its future have a source of information that almost certainly makes things worse rather than better: they can taste the wine. In addition, of course, even if they have a

good understanding of the effects of the weather on wine quality, they will not be able to maintain the consistency of a formula.

The most important development in the field since Meehl's original work is Robyn Dawes's famous article "The Robust Beauty of Improper Linear Models in Decision Making." The dominant statistical practice in the social sciences is to assign weights to the different predictors by following an algorithm, called multiple regression, that is now built into conventional software. The logic of multiple regression is unassailable: it finds the optimal formula for putting together a weighted combination of the predictors. However, Dawes observed that the complex statistical algorithm adds little or no value. One can do just as well by selecting a set of scores that have some validity for predicting the outcome and adjusting the values to make them comparable (by using standard scores or ranks). A formula that combines these predictors with equal weights is likely to be just as accurate in predicting new cases as the multiple-regression formula that was optimal in the original sample. More recent research went further: formulas that assign equal weights to all the predictors are often superior, because they are not affected by accidents of sampling.

The surprising success of equal-weighting schemes has an important practical implication: it is possible to develop useful algorithms without any prior statistical research. Simple equally weighted formulas based on existing statistics or on common sense are often very good predictors of significant outcomes. In a memorable example, Dawes showed that marital stability is well predicted by a formula:

frequency of lovemaking minus frequency of quarrels

You don't want your result to be a negative number.

The important conclusion from this research is that an algorithm that is constructed on the back of an envelope is often good enough to compete with an optimally weighted formula, and certainly good enough to outdo expert judgment. This logic can be applied in many domains, ranging from the selection of stocks by portfolio managers to the choices of medical treatments by doctors or patients.

A classic application of this approach is a simple algorithm that has saved the lives of hundreds of thousands of infants. Obstetricians had always known that an infant who is not breathing normally within a few

minutes of birth is at high risk of brain damage or death. Until the anesthesiologist Virginia Apgar intervened in 1953, physicians and midwives used their clinical judgment to determine whether a baby was in distress. Different practitioners focused on different cues. Some watched for breathing problems while others monitored how soon the baby cried. Without a standardized procedure, danger signs were often missed, and many newborn infants died.

One day over breakfast, a medical resident asked how Dr. Apgar would make a systematic assessment of a newborn. "That's easy," she replied. "You would do it like this." Apgar jotted down five variables (heart rate, respiration, reflex, muscle tone, and color) and three scores (0, 1, or 2, depending on the robustness of each sign). Realizing that she might have made a breakthrough that any delivery room could implement, Apgar began rating infants by this rule one minute after they were born. A baby with a total score of 8 or above was likely to be pink, squirming, crying, grimacing, with a pulse of 100 or more—in good shape. A baby with a score of 4 or below was probably bluish, flaccid, passive, with a slow or weak pulse—in need of immediate intervention. Applying Apgar's score, the staff in delivery rooms finally had consistent standards for determining which babies were in trouble, and the formula is credited for an important contribution to reducing infant mortality. The Apgar test is still used every day in every delivery room. Atul Gawande's recent *A Checklist Manifesto* provides many other examples of the virtues of checklists and simple rules.

THE HOSTILITY TO ALGORITHMS

From the very outset, clinical psychologists responded to Meehl's ideas with hostility and disbelief. Clearly, they were in the grip of an illusion of skill in terms of their ability to make long-term predictions. On reflection, it is easy to see how the illusion came about and easy to sympathize with the clinicians' rejection of Meehl's research.

The statistical evidence of clinical inferiority contradicts clinicians' everyday experience of the quality of their judgments. Psychologists who work with patients have many hunches during each therapy session, anticipating how the patient will respond to an intervention, guessing what will happen next. Many of these hunches are confirmed, illustrating the reality of clinical skill.

The problem is that the correct judgments involve short-term predictions in the context of the therapeutic interview, a skill in which therapists may have years of practice. The tasks at which they fail typically require long-term predictions about the patient's future. These are much more difficult, even the best formulas do only modestly well, and they are also tasks that the clinicians have never had the opportunity to learn properly—they would have to wait years for feedback, instead of receiving the instantaneous feedback of the clinical session. However, the line between what clinicians can do well and what they cannot do at all well is not obvious, and certainly not obvious to them. They know they are skilled, but they don't necessarily know the boundaries of their skill. Not surprisingly, then, the idea that a mechanical combination of a few variables could outperform the subtle complexity of human judgment strikes experienced clinicians as obviously wrong.

The debate about the virtues of clinical and statistical prediction has always had a moral dimension. The statistical method, Meehl wrote, was criticized by experienced clinicians as "mechanical, atomistic, additive, cut and dried, artificial, unreal, arbitrary, incomplete, dead, pedantic, fractionated, trivial, forced, static, superficial, rigid, sterile, academic, pseudoscientific and blind." The clinical method, on the other hand, was lauded by its proponents as "dynamic, global, meaningful, holistic, subtle, sympathetic, configural, patterned, organized, rich, deep, genuine, sensitive, sophisticated, real, living, concrete, natural, true to life, and understanding."

This is an attitude we can all recognize. When a human competes with a machine, whether it is John Henry a-hammerin' on the mountain or the chess genius Garry Kasparov facing off against the computer Deep Blue, our sympathies lie with our fellow human. The aversion to algorithms making decisions that affect humans is rooted in the strong preference that many people have for the natural over the synthetic or artificial. Asked whether they would rather eat an organic or a commercially grown apple, most people prefer the "all natural" one. Even after being informed that the two apples taste the same, have identical nutritional value, and are equally healthful, a majority still prefer the organic fruit. Even the producers of beer have found that they can increase sales by putting "All Natural" or "No Preservatives" on the label.

The deep resistance to the demystification of expertise is illustrated by the reaction of the European wine community to Ashenfelter's formula for predicting the price of Bordeaux wines. Ashenfelter's formula answered a prayer: one might thus have expected that wine lovers everywhere would be grateful to him for demonstrably improving their ability to identify the

wines that later would be good. Not so. The response in French wine circles, wrote *The New York Times*, ranged "somewhere between violent and hysterical." Ashenfelter reports that one oenophile called his findings "ludicrous and absurd." Another scoffed, "It is like judging movies without actually seeing them."

The prejudice against algorithms is magnified when the decisions are consequential. Meehl remarked, "I do not quite know how to alleviate the horror some clinicians seem to experience when they envisage a treatable case being denied treatment because a 'blind, mechanical' equation misclassifies him." In contrast, Meehl and other proponents of algorithms have argued strongly that it is unethical to rely on intuitive judgments for important decisions if an algorithm is available that will make fewer mistakes. Their rational argument is compelling, but it runs against a stubborn psychological reality: for most people, the cause of a mistake matters. The story of a child dying because an algorithm made a mistake is more poignant than the story of the same tragedy occurring as a result of human error, and the difference in emotional intensity is readily translated into a moral preference.

Fortunately, the hostility to algorithms will probably soften as their role in everyday life continues to expand. Looking for books or music we might enjoy, we appreciate recommendations generated by software. We take it for granted that decisions about credit limits are made without the direct intervention of any human judgment. We are increasingly exposed to guidelines that have the form of simple algorithms, such as the ratio of good and bad cholesterol levels we should strive to attain. The public is now well aware that formulas may do better than humans in some critical decisions in the world of sports: how much a professional team should pay for particular rookie players, or when to punt on fourth down. The expanding list of tasks that are assigned to algorithms should eventually reduce the discomfort that most people feel when they first encounter the pattern of results that Meehl described in his disturbing little book.

LEARNING FROM MEEHL

In 1955, as a twenty-one-year-old lieutenant in the Israeli Defense Forces, I was assigned to set up an interview system for the entire army. If you wonder why such a responsibility would be forced upon someone so young, bear in mind that the state of Israel itself was only seven years old at the time; all its institutions were under construction, and someone had to build them. Odd

as it sounds today, my bachelor's degree in psychology probably qualified me as the best-trained psychologist in the army. My direct supervisor, a brilliant researcher, had a degree in chemistry.

An interview routine was already in place when I was given my mission. Every soldier drafted into the army completed a battery of psychometric tests, and each man considered for combat duty was interviewed for an assessment of personality. The goal was to assign the recruit a score of general fitness for combat and to find the best match of his personality among various branches: infantry, artillery, armor, and so on. The interviewers were themselves young draftees, selected for this assignment by virtue of their high intelligence and interest in dealing with people. Most were women, who were at the time exempt from combat duty. Trained for a few weeks in how to conduct a fifteen- to twenty-minute interview, they were encouraged to cover a range of topics and to form a general impression of how well the recruit would do in the army.

Unfortunately, follow-up evaluations had already indicated that this interview procedure was almost useless for predicting the future success of recruits. I was instructed to design an interview that would be more useful but would not take more time. I was also told to try out the new interview and to evaluate its accuracy. From the perspective of a serious professional, I was no more qualified for the task than I was to build a bridge across the Amazon.

Fortunately, I had read Paul Meehl's "little book," which had appeared just a year earlier. I was convinced by his argument that simple, statistical rules are superior to intuitive "clinical" judgments. I concluded that the then current interview had failed at least in part because it allowed the interviewers to do what they found most interesting, which was to learn about the dynamics of the interviewee's mental life. Instead, we should use the limited time at our disposal to obtain as much specific information as possible about the interviewee's life in his normal environment. Another lesson I learned from Meehl was that we should abandon the procedure in which the interviewers' global evaluations of the recruit determined the final decision. Meehl's book suggested that such evaluations should not be trusted and that statistical summaries of separately evaluated attributes would achieve higher validity.

I decided on a procedure in which the interviewers would evaluate several relevant personality traits and score each separately. The final score of fitness for combat duty would be computed according to a standard formula, with no further input from the interviewers. I made up a list of six charac-

teristics that appeared relevant to performance in a combat unit, including "responsibility," "sociability," and "masculine pride." I then composed, for each trait, a series of factual questions about the individual's life before his enlistment, including the number of different jobs he had held, how regular and punctual he had been in his work or studies, the frequency of his interactions with friends, and his interest and participation in sports, among others. The idea was to evaluate as objectively as possible how well the recruit had done on each dimension.

By focusing on standardized, factual questions, I hoped to combat the halo effect, where favorable first impressions influence later judgments. As a further precaution against halos, I instructed the interviewers to go through the six traits in a fixed sequence, rating each trait on a five-point scale before going on to the next. And that was that. I informed the interviewers that they need not concern themselves with the recruit's future adjustment to the military. Their only task was to elicit relevant facts about his past and to use that information to score each personality dimension. "Your function is to provide reliable measurements," I told them. "Leave the predictive validity to me," by which I meant the formula that I was going to devise to combine their specific ratings.

The interviewers came close to mutiny. These bright young people were displeased to be ordered, by someone hardly older than themselves, to switch off their intuition and focus entirely on boring factual questions. One of them complained, "You are turning us into robots!" So I compromised. "Carry out the interview exactly as instructed," I told them, "and when you are done, have your wish: close your eyes, try to imagine the recruit as a soldier, and assign him a score on a scale of 1 to 5."

Several hundred interviews were conducted by this new method, and a few months later we collected evaluations of the soldiers' performance from the commanding officers of the units to which they had been assigned. The results made us happy. As Meehl's book had suggested, the new interview procedure was a substantial improvement over the old one. The sum of our six ratings predicted soldiers' performance much more accurately than the global evaluations of the previous interviewing method, although far from perfectly. We had progressed from "completely useless" to "moderately useful."

The big surprise to me was that the intuitive judgment that the interviewers summoned up in the "close your eyes" exercise also did very well, indeed just as well as the sum of the six specific ratings. I learned from this finding a lesson that I have never forgotten: intuition adds value even in the

justly derided selection interview, but only after a disciplined collection of objective information and disciplined scoring of separate traits. I set a formula that gave the "close your eyes" evaluation the same weight as the sum of the six trait ratings. A more general lesson that I learned from this episode was do not simply trust intuitive judgment—your own or that of others—but do not dismiss it, either.

Some forty-five years later, after I won a Nobel Prize in economics, I was for a short time a minor celebrity in Israel. On one of my visits, someone had the idea of escorting me around my old army base, which still housed the unit that interviews new recruits. I was introduced to the commanding officer of the Psychological Unit, and she described their current interviewing practices, which had not changed much from the system I had designed; there was, it turned out, a considerable amount of research indicating that the interviews still worked well. As she came to the end of her description of how the interviews are conducted, the officer added, "And then we tell them, 'Close your eyes.'"

DO IT YOURSELF

The message of this chapter is readily applicable to tasks other than making manpower decisions for an army. Implementing interview procedures in the spirit of Meehl and Dawes requires relatively little effort but substantial discipline. Suppose that you need to hire a sales representative for your firm. If you are serious about hiring the best possible person for the job, this is what you should do. First, select a few traits that are prerequisites for success in this position (technical proficiency, engaging personality, reliability, and so on). Don't overdo it—six dimensions is a good number. The traits you choose should be as independent as possible from each other, and you should feel that you can assess them reliably by asking a few factual questions. Next, make a list of those questions for each trait and think about how you will score it, say on a 1–5 scale. You should have an idea of what you will call "very weak" or "very strong."

These preparations should take you half an hour or so, a small investment that can make a significant difference in the quality of the people you hire. To avoid halo effects, you must collect the information on one trait at a time, scoring each before you move on to the next one. Do not skip around. To evaluate each candidate, add up the six scores. Because you are in charge of the final decision, you should not do a "close your eyes." Firmly resolve that you will hire the candidate whose final score is the highest, even

if there is another one whom you like better—try to resist your wish to invent broken legs to change the ranking. A vast amount of research offers a promise: you are much more likely to find the best candidate if you use this procedure than if you do what people normally do in such situations, which is to go into the interview unprepared and to make choices by an overall intuitive judgment such as "I looked into his eyes and liked what I saw."

SPEAKING OF JUDGES VS. FORMULAS

"Whenever we can replace human judgment by a formula, we should at least consider it."

"He thinks his judgments are complex and subtle, but a simple combination of scores could probably do better."

"Let's decide in advance what weight to give to the data we have on the candidates' past performance. Otherwise we will give too much weight to our impression from the interviews."