Taming Intuitive Predictions

Life presents us with many occasions to forecast. Economists forecast inflation and unemployment, financial analysts forecast earnings, military experts predict casualties, venture capitalists assess profitability, publishers and producers predict audiences, contractors estimate the time required to complete projects, chefs anticipate the demand for the dishes on their menu, engineers estimate the amount of concrete needed for a building, fireground commanders assess the number of trucks that will be needed to put out a fire. In our private lives, we forecast our spouse's reaction to a proposed move or our own future adjustment to a new job.

Some predictive judgments, such as those made by engineers, rely largely on look-up tables, precise calculations, and explicit analyses of outcomes observed on similar occasions. Others involve intuition and System 1, in two main varieties. Some intuitions draw primarily on skill and expertise acquired by repeated experience. The rapid and automatic judgments and choices of chess masters, fireground commanders, and physicians that Gary Klein has described in *Sources of Power* and elsewhere illustrate these skilled intuitions, in which a solution to the current problem comes to mind quickly because familiar cues are recognized.

Other intuitions, which are sometimes subjectively indistinguishable from the first, arise from the operation of heuristics that often substitute an easy question for the harder one that was asked. Intuitive judgments can be made with high confidence even when they are based on nonregressive assessments of weak evidence. Of course, many judgments, especially in the professional domain, are influenced by a combination of analysis and intuition.

Nonregressive Intuitions

Let us return to a person we have already met:

> Julie is currently a senior in a state university. She read fluently when she was four years old. What is her grade point average (GPA)?

People who are familiar with the American educational scene quickly come up with a number, which is often in the vicinity of 3.7 or 3.8. How does this occur? Several operations of System 1 are involved.

- A causal link between the evidence (Julie's reading) and the target of the prediction (her GPA) is sought. The link can be indirect. In this instance, early reading and a high GDP are both indications of academic talent. Some connection is necessary. You (your System 2) would probably reject as irrelevant a report of Julie winning a fly fishing competitiowhired D=n or excelling at weight lifting in high school. The process is effectively dichotomous. We are capable of rejecting information as irrelevant or false, but adjusting for smaller weaknesses in the evidence is not something that System 1 can do. As a result, intuitive predictions are almost completely insensitive to the actual predictive quality of the evidence. When a link is found, as in the case of Julie's early reading, WY SIATI applies: your associative memory quickly and automatically constructs the best possible story from the information available.
- Next, the evidence is evaluated in relation to a relevant norm. How precocious is a child who reads fluently at age four? What relative rank or percentile score corresponds to this achievement? The group to which the child is compared (we call it a reference group) is not fully specified, but this is also the rule in normal speech: if someone graduating from college is described as "quite clever" you rarely need to ask, "When you say 'quite clever,' which reference group do you have in mind?"
- The next step involves substitution and intensity matching. The evaluation of the flimsy evidence of cognitive ability in childhood is substituted as an answer to the question about her college GPA. Julie will be assigned the same percentile score for her GPA and for her achievements as an early reader.
- The question specified that the answer must be on the GPA scale, which requires another intensity-matching operation, from a general impression of Julie's academic achievements to the GPA that matches the evidence for her talent. The final step is a translation, from an impression of Julie's relative academic standing to the GPA that corresponds to it.

Intensity matching yields predictions that are as extreme as the evidence on which they are based, leading people to give the same answer to two quite different questions:

> What is Julie's percentile score on reading precocity?
> What is Julie's percentile score on GPA?

By now you should easily recognize that all these operations are features of System 1. I listed them here as an orderly sequence of steps, but of course the spread of activation in associative memory does not work this way. You should imagine a process of spreading activation that is initially prompted by the evidence and the question, feeds back upon itself, and eventually settles on the most coherent solution possible.

Amos and I once asked participants in an experiment to judge descriptions of eight college freshmen, allegedly written by a counselor on the basis of interviews of the entering class. Each description consisted of five adjectives, as in the following example:

> intelligent, self-confident, well-read, hardworking, inquisitive

We asked some participants to answer two questions:

> How much does this description impress you with respect to academic ability?
>
> What percentage of descriptions of freshmen do you believe would impress you more?

The questions require you to evaluate the evidence by comparing the description to your norm for descriptions of students by counselors. The very existence of such a norm is remarkable. Although you surely do not know how you acquired it, you have a fairly clear sense of how much enthusiasm the description conveys: the counselor believes that this student is good, but not spectacularly good. There is room for stronger adjectives than *intelligent* (*brilliant*, *creative*), *well-read* (*scholarly, erudite, impressively knowledgeable*), and *hardworking* (*passionate*, *perfectionist*). The verdict: very likely to be in the top 15% but unlikely to be in the top 3%. There is impressive consensus in such judgments, at least within a culture.

The other participants in our experiment were asked different questions:

> What is your estimate of the grade point average that the student will obtain?
> What is the percentage of freshmen who obtain a higher GPA?

You need another look to detect the subtle difference between the two

sets of questions. The difference should be obvious, but it is not. Unlike the first questions, which required you only to evaluate the evidence, the second set involves a great deal of uncertainty. The question refers to actual performance at the end of the freshman year. What happened during the year since the interview was performed? How accurately can you predict the student's actual achievements in the first year at college from five adjectives? Would the counselor herself be perfectly accurate if she predicted GPA from an interview?

The objective of this study was to compare the percentile judgments that the participants made when evaluating the evidence in one case, and when predicting the ultimate outcome in another. The results are easy to summarize: the judgments were identical. Although the two sets of questions differ (one is about the description, the other about the student's future academic performance), the participants treated them as if they were the same. As was the case with Julie, the prediction of the future is not distinguished from an evaluation of current evidence—prediction matches evaluation. This is perhaps the best evidence we have for the role of substitution. People are asked for a prediction but they substitute an evaluation of the evidence, without noticing that the question they answer is not the one they were asked. This process is guaranteed to generate predictions that are systematically biased; they completely ignore regression to the mean.

During my military service in the Israeli Defense Forces, I spent some time attached to a unit that selected candidates for officer training on the basis of a series of interviews and field tests. The designated criterion for successful prediction was a cadet's final grade in officer school. The validity of the ratings was known to be rather poor (I will tell more about it in a later chapter). The unit still existed years later, when I was a professor and collaborating with Amos in the study of intuitive judgment. I had good contacts with the people at the unit and asked them for a favor. In addition to the usual grading system they used to evaluate the candidates, I asked for their best guess of the grade that each of the future cadets would obtain in officer school. They collected a few hundred such forecasts. The officers who had produced the prediof pctions were all familiar with the letter grading system that the school applied to its cadets and the approximate proportions of A's, B's, etc., among them. The results were striking: the relative frequency of A's and B's in the predictions was almost identical to the frequencies in the final grades of the school.

These findings provide a compelling example of both substitution and intensity matching. The officers who provided the predictions completely failed to discriminate between two tasks:

- their usual mission, which was to evaluate the performance of candidates during their stay at the unit
- the task I had asked them to perform, which was an actual prediction of a future grade

They had simply translated their own grades onto the scale used in officer school, applying intensity matching. Once again, the failure to address the (considerable) uncertainty of their predictions had led them to predictions that were completely nonregressive.

A Correction for Intuitive Predictions

Back to Julie, our precocious reader. The correct way to predict her GPA was introduced in the preceding chapter. As I did there for golf on successive days and for weight and piano playing, I write a schematic formula for the factors that determine reading age and college grades:

> reading age = shared factors + factors specific to reading age = 100%
> GPA = shared factors + factors specific to GPA = 100%

The shared factors involve genetically determined aptitude, the degree to which the family supports academic interests, and anything else that would cause the same people to be precocious readers as children and academically successful as young adults. Of course there are many factors that would affect one of these outcomes and not the other. Julie could have been pushed to read early by overly ambitious parents, she may have had an unhappy love affair that depressed her college grades, she could have had a skiing accident during adolescence that left her slightly impaired, and so on.

Recall that the correlation between two measures—in the present case reading age and GPA—is equal to the proportion of shared factors among their determinants. What is your best guess about that proportion? My most optimistic guess is about 30%. Assuming this estimate, we have all we need to produce an unbiased prediction. Here are the directions for how to get there in four simple steps:

1. Start with an estimate of average GPA.
2. Determine the GPA that matches your impression of the evidence.
3. Estimate the correlation between your evidence and GPA.
4. If the correlation is .30, move 30% of the distance from the average to the matching GPA.

Step 1 gets you the baseline, the GPA you would have predicted if you were told nothing about Julie beyond the fact that she is a graduating senior. In the absence of information, you would have predicted the average. (This is similar to assigning the base-rate probability of business administration grahavpduates when you are told nothing about Tom W.) Step 2 is your intuitive prediction, which matches your evaluation of the evidence. Step 3 moves you from the baseline toward your intuition, but the distance you are allowed to move depends on your estimate of the correlation. You end up, at step 4, with a prediction that is influenced by your intuition but is far more moderate.

This approach to prediction is general. You can apply it whenever you need to predict a quantitative variable, such as GPA, profit from an investment, or the growth of a company. The approach builds on your intuition, but it moderates it, regresses it toward the mean. When you have good reasons to trust the accuracy of your intuitive prediction—a strong correlation between the evidence and the prediction—the adjustment will be small.

Intuitive predictions need to be corrected because they are not regressive and therefore are biased. Suppose that I predict for each golfer in a tournament that his score on day 2 will be the same as his score on day 1. This prediction does not allow for regression to the mean: the golfers who fared well on day 1 will on average do less well on day 2, and those who did poorly will mostly improve. When they are eventually compared to actual outcomes, nonregressive predictions will be found to be biased. They are on average overly optimistic for those who did best on the first day and overly pessimistic for those who had a bad start. The predictions are as extreme as the evidence. Similarly, if you use childhood achievements to predict grades in college without regressing your predictions toward the mean, you will more often than not be disappointed by the academic outcomes of early readers and happily surprised by the grades of those who learned to read relatively late. The corrected intuitive predictions eliminate these biases, so that predictions (both high and low) are about equally likely to overestimate and to underestimate the true value. You still make errors when your predictions are unbiased, but the errors are smaller and do not favor either high or low outcomes.

A Defense of Extreme Predictions?

I introduced Tom W earlier to illustrate predictions of discrete outcomes such as field of specialization or success in an examination, which are expressed by assigning a probability to a specified event (or in that case by ranking outcomes from the most to the least probable). I also described a procedure that counters the common biases of discrete prediction: neglect of base rates and insensitivity to the quality of information.

The biases we find in predictions that are expressed on a scale, such as GPA or the revenue of a firm, are similar to the biases observed in judging the probabilities of outcomes.

The corrective procedures are also similar:

- Both contain a baseline prediction, which you would make if you knew nothing about the case at hand. In the categorical case, it was the base rate. In the numerical case, it is the average outcome in the relevant category.
- Both contain an intuitive prediction, which expresses the number that comes to your mind, whether it is a probability or a GPA.
- In both cases, you aim for a prediction that is intermediate between the baseline and your intuitive response.
- In the default case of no useful evidence, you stay with the baseline.
- At the other extreme, you also stay with your initial predictiononsp. This will happen, of course, only if you remain completely confident in your initial prediction after a critical review of the evidence that supports it.
- In most cases you will find some reason to doubt that the correlation between your intuitive judgment and the truth is perfect, and you will end up somewhere between the two poles.

This procedure is an approximation of the likely results of an appropriate statistical analysis. If successful, it will move you toward unbiased predictions, reasonable assessments of probability, and moderate predictions of numerical outcomes. The two procedures are intended to address the same bias: intuitive predictions tend to be overconfident and overly extreme.

Correcting your intuitive predictions is a task for System 2. Significant effort is required to find the relevant reference category, estimate the baseline prediction, and evaluate the quality of the evidence. The effort is justified only when the stakes are high and when you are particularly keen not to make mistakes. Furthermore, you should know that correcting your intuitions may complicate your life. A characteristic of unbiased predictions is that they permit the prediction of rare or extreme events only when the information is very good. If you expect your predictions to be of modest validity, you will never guess an outcome that is either rare or far from the mean. If your predictions are unbiased, you will never have the satisfying experience of correctly calling an extreme case. You will never be able to say, "I thought so!" when your best student in law school becomes a Supreme Court justice, or when a start-up that you thought very promising eventually becomes a major commercial success. Given the limitations of the evidence, you will never predict that an outstanding high school student will be a straight-A student at Princeton. For the same reason, a venture capitalist will never be told that the probability of success for a start-up in its early stages is "very high."

The objections to the principle of moderating intuitive predictions must be taken seriously, because absence of bias is not always what matters most. A preference for unbiased predictions is justified if all errors of prediction are treated alike, regardless of their direction. But there are situations in which one type of error is much worse than another. When a venture capitalist looks for "the next big thing," the risk of missing the next Google or Facebook is far more important than the risk of making a modest investment in a start-up that ultimately fails. The goal of venture capitalists is to call the extreme cases correctly, even at the cost of overestimating the prospects of many other ventures. For a conservative banker making large loans, the risk of a single borrower going bankrupt may outweigh the risk of turning down several would-be clients who would fulfill their obligations. In such cases, the use of extreme language ("very good prospect," "serious risk of default") may have some justification for the comfort it provides, even if the information on which these judgments are based is of only modest validity.

For a rational person, predictions that are unbiased and moderate should not present a problem. After all, the rational venture capitalist knows that even the most promising start-ups have only a moderate chance of success. She views her job as picking the most promising bets from the bets that are available and does not feel the need to delude herself about the prospects of a start-up in which she plans to invest. Similarly, rational individuals predicting the revenue of a firm will not be bound to a singleys p number—they should consider the range of uncertainty around the most

likely outcome. A rational person will invest a large sum in an enterprise that is most likely to fail if the rewards of success are large enough, without deluding herself about the chances of success. However, we are not all rational, and some of us may need the security of distorted estimates to avoid paralysis. If you choose to delude yourself by accepting extreme predictions, however, you will do well to remain aware of your self-indulgence.

Perhaps the most valuable contribution of the corrective procedures I propose is that they will require you to think about how much you know. I will use an example that is familiar in the academic world, but the analogies to other spheres of life are immediate. A department is about to hire a young professor and wants to choose the one whose prospects for scientific productivity are the best. The search committee has narrowed down the choice to two candidates:

> Kim recently completed her graduate work. Her recommendations are spectacular and she gave a brilliant talk and impressed everyone in her interviews. She has no substantial track record of scientific productivity.

> Jane has held a postdoctoral position for the last three years. She has been very productive and her research record is excellent, but her talk and interviews were less sparkling than Kim's.

The intuitive choice favors Kim, because she left a stronger impression, and WYSIATI. But it is also the case that there is much less information about Kim than about Jane. We are back to the law of small numbers. In effect, you have a smaller sample of information from Kim than from Jane, and extreme outcomes are much more likely to be observed in small samples. There is more luck in the outcomes of small samples, and you should therefore regress your prediction more deeply toward the mean in your prediction of Kim's future performance. When you allow for the fact that Kim is likely to regress more than Jane, you might end up selecting Jane although you were less impressed by her. In the context of academic choices, I would vote for Jane, but it would be a struggle to overcome my intuitive impression that Kim is more promising. Following our intuitions is more natural, and somehow more pleasant, than acting against them.

You can readily imagine similar problems in different contexts, such as a venture capitalist choosing between investments in two start-ups that operate in different markets. One start-up has a product for which demand

can be estimated with fair precision. The other candidate is more exciting and intuitively promising, but its prospects are less certain. Whether the best guess about the prospects of the second start-up is still superior when the uncertainty is factored in is a question that deserves careful consideration.

A Two-Systems View of Regression

Extreme predictions and a willingness to predict rare events from weak evidence are both manifestations of System 1. It is natural for the associative machinery to match the extremeness of predictions to the perceived extremeness of evidence on which it is based—this is how substitution works. And it is natural for System 1 to generate overconfident judgments, because confidence, as we have seen, is determined by the coherence of the best story you can tell from the evidence at hand. Be warned: your intuitions will deliver predictions that are too extreme and you will be inclinehe pd to put far too much faith in them.

Regression is also a problem for System 2. The very idea of regression to the mean is alien and difficult to communicate and comprehend. Galton had a hard time before he understood it. Many statistics teachers dread the class in which the topic comes up, and their students often end up with only a vague understanding of this crucial concept. This is a case where System 2 requires special training. Matching predictions to the evidence is not only something we do intuitively; it also seems a reasonable thing to do. We will not learn to understand regression from experience. Even when a regression is identified, as we saw in the story of the flight instructors, it will be given a causal interpretation that is almost always wrong.

Speaking of Intuitive Predictions

"That start-up achieved an outstanding proof of concept, but we shouldn't expect them to do as well in the future. They are still a long way from the market and there is a lot of room for regression."

"Our intuitive prediction is very favorable, but it is probably too high. Let's take into account the strength of our evidence and regress the prediction toward the mean."

“The investment may be a good idea, even if the best guess is that it will fail. Let's not say we really believe it is the next Google.”

“I read one review of that brand and it was excellent. Still, that could have been a fluke. Let’s consider only the brands that have a large number of reviews and pick the one that looks best.”