5º

book club synopsis

october 2015

Superforecasting: The Art and Science of Prediction
Philip E. Tetlock and Dan Gardner

Crown, 2015

Superforecasting: The Art and Science of Prediction is a book about forecasting, what makes some people particularly good at it and how we can learn from their examples in order to become better at making accurate predictions.

The book draws on decades of research and the results of a large government-funded forecasting tournament. Much of *Superforecasting* is devoted to the results of the Good Judgment Project, which involved more than 20,000 volunteers from all walks of life who set out to forecast global political and economic events.

About the authors

Philip E. Tetlock is the Annenberg University Professor at the University of Pennsylvania and holds appointments in the Psychology and Political Science departments and the Wharton School of Business. He and his wife, Barbara Mellers, are co-leaders of the Good Judgment Project, a multiyear forecasting study.

Dan Gardner is a journalist and the author of *Risk* and *Future Babble: Why Pundits are Hedgehogs and Foxes Know Best.*

contents:

chapter one:

an optimistic skeptic

The opening chapter begins with a comparison of two forecasters: Tom Friedman and Bill Flack.

Friedman is a *New York Times* political columnist and the author of several best-selling books. He appears regularly on CNN and is a frequent guest at Davos. Millions of people look to him to tell them not only what's happening now and why it's happening, but what will happen in the future.

Then there's Flack, a retired Department of Agriculture scientist from Nebraska. He also forecasts global events, but no one is asking Flack to appear on CNN, and he's never been invited to sit on a panel at Davos. That's unfortunate, Tetlock says, because he's "a remarkable forecaster."

"We know that because each one of Bill's predictions has been dated, recorded and assessed for accuracy by independent scientific observers. His track record is excellent," writes Tetlock.

Among the 300 questions that Flack has answered accurately are:

- Will Russia officially annex additional Ukrainian territory in the next three months?
- Will North Korea detonate a nuclear device before the end of the year?
- Will India or Brazil become a permanent member of the UN Security Council in the next two years?

Flack is what Tetlock calls a "superforecaster," someone with a seemingly uncanny ability to accurately predict the likelihood of future events. It's a distinction he shares with about 2 percent of the 20,000 forecasters who volunteered to take part in a multiyear forecasting study lead by Tetlock and his research partners.

Forecasting, Tetlock says, "is not a 'you have it or you don't' talent. It is a skill that can be cultivated. This book will show you how."

Tetlock's earlier landmark study

From 1984 to 2004, Tetlock conducted "the most comprehensive assessment of expert judgment in the scientific literature." He gathered a large group of experts and asked them to make thousands of predictions about the economy, stocks, elections, wars and other economic and political issues.

The main conclusion of the study was that these experts did about as well as random guessing, or as Tetlock famously said, "the average expert was roughly as accurate as a dart-throwing chimpanzee."

That study and the book that followed, *Expert Political Judgment*, garnered a huge amount of media attention. But he writes, "I realized that as word of my work spread, its apparent meaning was mutating." His research had shown that the average expert had done little better than guessing on many of the political and economic questions that he had posed. But, he adds, "'many' does not equal all."

"The message became 'all expert forecasts are useless,' which is nonsense," he writes. He says his research didn't support these more extreme conclusions, and today that is all the more true. He believes it is possible to see into the future, at least somewhat, "and that any intelligent, open-minded and hardworking person can cultivate the requisite skills." Regarding forecasting, he considers himself an "optimistic skeptic."

The story of Mohamed Bouazizi

To explain the "skeptic" half of that label, Tetlock recounts the story of Mohamed Bouazizi, a 26-year-old Tunisian man who made

his living selling fruits and vegetables from a wooden handcart. To support his family, Bouazizi had to borrow money to fill a cart with produce and then hope to sell enough produce to be able to pay off the debt and have a little left over.

One morning, the police stopped him and told him they were taking away his scales because he had violated a regulation. Bouazizi knew the police were lying; he had not violated any regulations. A policewoman slapped and insulted him, and the authorities took away his scales and his produce.

When Bouazizi went to a town office to complain, he was told the official was busy in a meeting. Bouazizi left the office and returned with fuel. Outside the office, he doused himself, lit a match and began to burn.

Bouazizi's self-immolation sparked protests in Tunisia, eventually leading the Tunisian president to visit him in the hospital. Bouazizi died on January 4, 2011, and 10 days later, the president fled to Saudi Arabia, ending a reign that lasted more than 20 years.

Soon afterward, protests erupted throughout the Arab world in what became known as the Arab Spring. And it all started with one poor man being harassed by the police. But, Tetlock says, while it's one thing to look back at this chain of events and sketch a narrative arc connecting Bouazizi to all the events that flowed from his protest, not even Friedman—an expert at piecing together such narratives in hindsight—could have looked into the future and foreseen all that followed.

The flap of a butterfly's wings

To further illustrate the limits on predictability, Tetlock discusses a 1972 paper by an American meteorologist named Edward Lorenz entitled, "Predictability: Does the Flap of a Butterfly's Wings in Brazil Set Off a Tornado in Texas?"

Ten years prior, Lorenz had discovered that tiny data entry variations in computer simulations of weather patterns could produce

dramatically different long-term forecasts. So, in principle, a lone butterfly in Brazil could flap its wings and set off a tornado in Texas.

"Of course Lorenz didn't mean that the butterfly 'causes' the tornado in the same sense that I cause a wine glass to break when I hit it with a hammer," writes Tetlock. "He meant that if that particular butterfly hadn't flapped its wings at that moment, the unfathomably complex network of atmospheric actions and reactions would have behaved differently, and the tornado might never have formed."

The optimist

The majority of chapter one is devoted to why Tetlock remains an optimist about the accuracy of forecasting, despite these two previous case studies.

He starts off by making the point that much of day-to-day living involves making accurate predictions—from when we leave for work in the morning to how we drive on the highway and a host of other mundane predictions.

But these "pockets of predictability" can be abruptly punctured. For example, you might plan to go to a restaurant that is normally open for dinner, but it may not open when you think it will, due to any number of reasons: from a manager sleeping late to bankruptcy or "a physics experiment accidentally creating a black hole that sucks up the solar system."

How predictable something is, Tetlock says, "depends on what we are trying to predict, how far into the future and under what circumstances."

Meteorologists know better than anyone that the further we try to look into the future, the harder it is to see. That's why one- and two-day forecasts are typically quite accurate, while eight-day forecasts are not.

The problem is that forecasters don't measure their accuracy. "Accuracy is seldom determined after the fact and is almost never

done with sufficient regularity and rigor that conclusions can be drawn," Tetlock writes. "The reason? Mostly it's a demand-side problem: The consumers of forecasting—governments, business and the public—don't demand evidence of accuracy. So there is no measurement. Which means no revision. And without revision, there can be no improvement."

That's where the Good Judgment Project comes in.

The Good Judgment Project

In the summer of 2011, Tetlock and his research partner Barbara Mellers (who is also his wife) launched the Good Judgment Project, inviting volunteers to make economic and political predictions about the future. Flack and about 2,000 others signed up for the project the first year; that number rose to more than 20,000 over the next four years.

The Good Judgment Project began as part of a much larger research effort sponsored by the Intelligence Advanced Research Projects Activity (IARPA), an agency within the intelligence community that reports to the director of national intelligence.

The job of IARPA is to support daring research that promises to make American intelligence better at what it does. Although forecasting is critical to national security, little was known about the quality of these forecasts, which is why IARPA decided to create a forecasting tournament and invite five scientific teams led by top researchers to compete to generate accurate forecasts on the kinds of tough questions that intelligence analysts deal with every day.

The Good Judgment Project was one of those five teams. Each team could use its own methods, but all five teams were required to submit forecasts for the same questions at the same time every day, from September 2011 to May 2015.

Over four years, IARPA posed nearly 5,000 questions about world affairs. Unlike Tetlock's earlier research, the time frames of the forecasts were shorter, with most extending out more than one month and less than one year.

In the first year of the tournament, the Good Judgment Project beat the official control group by 60 percent. In year two, it beat the control group by 78 percent. The Good Judgment Project even outperformed professional intelligence analysts who had access to classified data. After two years, the Good Judgment Project was doing so much better than all of the other teams that IARPA dropped the other teams.

Tetlock says two key conclusions emerged from the research:

1. Foresight is real. Superforecasters "have a real, measurable skill at judging how high-stakes events are likely to unfold three months, six months, a year, or a year and a half in advance."
2. It's not really who these superforecasters are that makes them so good; it's what they do. "Foresight isn't a mysterious gift bestowed at birth," Tetlock writes. "It is the product of particular ways of thinking, of gathering information, of updating beliefs."

He says these habits of thought "can be learned and cultivated by any intelligent, thoughtful, determined person."

One surprising result was the fact that a tutorial covering some basic concepts that took about an hour to read improved accuracy by about 10 percent through the entire tournament year.

But what is it that elevates forecasting to superforecasting? What matters most, says Tetlock, is how the forecaster thinks. Broadly speaking, superforecasting "demands thinking that is open-minded, careful, curious, and—above all—self-critical."

Superforecasting also demands focus. "The kind of thinking that produces superior judgment does not come effortlessly," he writes.

A forecast about forecasting

The chapter ends with a discussion of whether the subjective judgment involved in superforecasting will even be necessary in the future, given that we live "in an era of dazzling powerful computers, incomprehensible algorithms and Big Data."

In 1954, a psychologist named Paul Meehl wrote a book that caused a big stir. He reviewed 20 studies that showed that well-informed experts predicting outcomes such as whether a student would succeed in college or a parolee would be sent back to prison were not as accurate as simple algorithms that add up objective indicators like ability test scores and records of past conduct.

Since then, more than 200 studies have proven that in most cases, statistical algorithms beat subjective judgment. What's more, advances in information technology suggest "we are approaching a historical discontinuity in humanity's relationship with machines." Tetlock cites IBM's Deep Blue beating Garry Kasparov in chess in 1997 and IBM's Watson beating two *Jeopardy!* champions in 2011 as examples.

So are we nearing the day when a supercomputer like Watson will be able to answer the question: "Will two top Russian leaders trade jobs in the next ten years?"

Tetlock spoke to Watson's chief engineer, David Ferrucci, who says he isn't sure a computer will ever be able to master subjective judgment. According to Ferrucci, machines may get better at mimicking human meaning and thereby predicting human behavior, but "there's a difference between mimicking and reflecting and originating meaning."

At the same time, Tetlock believes that in forecasting, as in other fields, we'll continue to see more examples of humans and computers competing together as teams. "To reframe the man-versus-machine dichotomy, combinations of Garry Kasparov and Deep Blue may prove more robust than pure-human or pure-machine approaches."

Both Tetlock and Ferrucci also believe that in the future, people will no longer listen to the advice of pundits whose views are informed solely by their subjective judgment. Instead, human experts using subjective judgment will be paired with computers in order to overcome the limitations and biases of human cognition.

chapter two:

illusions of knowledge

In chapter one, Tetlock asks, "If it's possible to improve foresight simply by measuring, and if the rewards of improved foresight are substantial, why isn't measuring standard practice?"

The author believes a big part of that answer has to do with the psychology that convinces us that we know things when we really don't. Chapter two is devoted to exploring that psychology.

To illustrate this, he begins the chapter with the story of Archie Cochrane, a physician who was diagnosed with basal cell carcinoma. After the skin cancer was removed, Cochrane's dermatologist sent him to see another specialist as a precaution. The specialist discovered a lump in Cochrane's armpit and recommended that it be surgically removed.

Cochrane awoke from that surgery to find that his whole chest was wrapped in bandages and the specialist was standing at his bedside with a grim expression. The specialist told Cochrane that his armpit was full of cancerous tissue and that although he had done his best to remove as much cancer as he could, "I may well not have saved your life."

Cochrane was devastated, and the next day started working out a plan for how he would spend his remaining time. But it turns out that the specialist was wrong. Cochrane didn't have terminal cancer. In fact, he didn't have cancer at all, which a pathologist confirmed after examining the tissue that was removed.

"Cochrane didn't doubt the specialist, and the specialist didn't doubt his own judgment, so neither man considered the possibility that the diagnosis was wrong and neither thought it wise to wait for the pathologist's report before closing the books on the life of Archie Cochrane," Tetlock writes.

That's human nature, the author says. We are all too quick to make up our minds and too slow to change them.

The history of medicine

As Tetlock notes, the "long and wretched history of medicine" provides plenty of evidence for this kind of faulty thinking.

Galen, the second-century physician to Roman emperors, never conducted anything resembling a modern scientific experiment, and yet he was steadfast in his conviction that his treatments were effective. "All those who drink of this treatment recover in a short time, except those whom it does not help, who all die," he wrote. "It is obvious, therefore, that it fails only in incurable cases."

Galen and others like him appear repeatedly in the history of medicine, Tetlock says. "They are men (always men) of strong conviction and a profound trust in their own judgment. They embrace treatments, develop bold theories for why they work, denounce rivals as quacks and charlatans and spread their insights with evangelical passion."

Putting medicine to the test

It wasn't until the 20th century that the ideas of randomized clinical trials, careful measurement and statistical power took hold. Doctors and scientists realized that if you wanted to test whether a treatment was effective, you needed to randomly assign people to either the treatment group or the nontreatment group. Randomly assigning people to one group or the other would mean that whatever differences there were among the individuals in the groups would balance out, and scientists could confidently conclude that the treatment caused any difference in outcomes.

But as Tetlock notes, "the idea of randomized controlled trials was painfully slow to catch on." Much of the medical establishment wasn't interested or was hostile to the idea. Archie Cochrane called this "the God complex." Doctors didn't need scientific validation. They just knew they were right.

Tetlock recounts the story of an early clinical trial that Cochrane proposed. He wanted to see if patients treated at newly created cardiac care units did better than patients who had the old treatment, which was to be sent home for monitoring and bed rest. Physicians were opposed to this idea. Most believed that it was so obvious that cardiac care units were superior that it would be unethical to deny all patients the best care. Cochrane persisted, however, and eventually the trial began.

Halfway through the trial, Cochrane met with the doctors who tried to block the study and shared the preliminary results with them. He told them that although the difference between the two treatments wasn't statistically significant, it appeared that the cardiac care unit group had slightly better care. The doctors were livid and demanded that Cochrane stop the trial at once. But then Cochrane revealed that he had played a trick on the doctors. The actual results were the opposite of what he had said. The group who had been sent home had done slightly better than the group in the cardiac care units. "The only alternative to a controlled experiment that delivers real insight is an uncontrolled experiment that produces merely the illusion of insight," says Tetlock, and that's exactly what these doctors had failed to grasp.

Thinking about thinking

Why are people so quick to rush to judgment without waiting for all the necessary information? Why do people make firm conclusions based on nothing more than intuition?

To explore the answers to these questions, Tetlock explains that modern psychologists describe how we think and decide by using a dual-system model that partitions our mental universe into two domains. "System 2 is the familiar realm of conscious thought. It

consists of everything we choose to focus on," Tetlock writes. "By contrast, System 1 is largely a stranger to us. It is the realm of automatic perceptual and cognitive operations—like those you are running right now to transform the print on this page into a meaningful sentence or to hold the book while reaching for a glass and taking a sip."

Tetlock says System 1 comes first and is constantly running in the background. If you are asked a question and you immediately know the answer, it springs from System 1. What System 2 does is interrogate that answer. Does it stand up to scrutiny? Is it backed by evidence? This process takes time and effort.

To illustrate this point, Tetlock asks this question: "A bat and a ball together cost $1.10. The bat costs a dollar more than the ball. How much does the ball cost?"

Using System 1, the answer that most people immediate come up with is 10 cents. But is that the right answer? Using System 2 to think about the question more carefully, it becomes clear that 10 cents is wrong. If the ball were to cost 10 cents, the bat would cost $1.10, bringing the total to $1.20. The correct answer is 5 cents.

The question shows that most people, even very intelligent people, aren't very reflective. We don't stop to think carefully and instead tend to go with strong and quick hunches. If it feels true, we assume it is true.

To illustrate our tendency to make snap judgments, Tetlock uses the story of a 2011 car bombing in Oslo, in which eight people were killed and more than 200 were injured. The immediate speculation on the Internet and cable news was that it had to be the work of radical Islamists. People rushed to Google to find evidence to support their hunch, and they succeeded. Norway has a poorly integrated Muslim community, and a radical Muslim preacher had been charged with incitement the week before.

Given the facts that were known at the time, it was reasonable to suspect Islamic terrorists. But the real perpetrator turned out to be Anders Breivik, a man who is not a Muslim, who in fact hates Muslims, writes Tetlock.

What scientists understand, Tetlock says, is that no matter how strongly they feel that they know the truth, they must also consider alternative explanations beyond hypotheses that are merely plausible. "[Scientists] must set that feeling aside and replace it with finely measured degrees of doubt—doubt that can be reduced (although never to zero) by better evidence from better studies."

This kind of scientific caution runs against human nature, Tetlock says. "Our natural inclination is to grab on to the first plausible explanation and happily gather supportive evidence without checking its reliability," he writes. "That is what psychologists call confirmation bias."

Bait and switch

Tetlock says there's another mental process at work when we make snap judgments without considering all the possibilities. The formal term is "attribute substitution," but Tetlock calls it "bait and switch."

"When faced with a hard question, we often surreptitiously replace it with an easy one," he writes. For example, Tetlock suspects that when Cochrane was told by the specialist that he had cancer, instead of asking himself, "Do I have cancer?" Cochrane substituted the question with, "Is this the sort of person who should know I have cancer?" And the answer to that second question was, "Of course! He is an eminent cancer specialist. He saw the cancerous flesh with his own eyes. This is *exactly* the sort of person who should know whether I have cancer."

Blinking and thinking

Tetlock calls our automatic, nearly effortless mode of thinking about the world the "tip-of-your-nose perspective." And although it's an imperfect view, Tetlock says we shouldn't discount it entirely. While

many books draw a dichotomy between intuition ("blink") and analysis ("think"), Tetlock believes it's a false dichotomy. The choice isn't about either/or, he says—it's about how to blend them in evolving situations.

To illustrate that point, he recounts a story told by a psychologist named Gary Klein about a firefighting commander who responded to a routine kitchen fire and ordered his men to stand in the living room and hose down the flames. Although the fire subsided at first, it quickly roared back. The commander was puzzled. He also noticed that the living room was surprisingly hot given the size of the kitchen fire.

The commander started feeling uneasy and quickly ordered everyone out of the house. Just as the firefighters reached the street, the floor in the living room collapsed because the real source of the fire was in the basement and not the kitchen. In this case, the commander didn't know how he knew to order everyone out of the house. It was an intuitive judgment, but it is nothing mystical, Tetlock writes. "It's pattern recognition. With training or experience, people can encode patterns deep in their memories in vast number and intricate detail. … If something doesn't fit a pattern—like a kitchen fire giving off more heat than a kitchen fire should—a competent expert senses it immediately."

He concludes the chapter with the point that while the tip-of-your-nose perspective can often work wonders, it's also important to consider the fact that what seems obviously true now may turn out to be false later.

chapter three:

keeping score

In this chapter, Tetlock examines what it takes to test forecasting as rigorously as modern medicine tests treatments.

He explains that although it seems fairly straightforward—collect forecasts, judge their accuracy and add the numbers—in reality, it's not nearly so simple to judge the accuracy of forecasts.

Take, for instance, the story of Microsoft CEO Steve Ballmer's infamous 2007 forecast about the iPhone: "There's no chance that the iPhone is going to get any significant market share. No chance."

While Ballmer's prediction is widely thought to be "enshrined in the forecasting hall of shame," Tetlock argues that if you parse Ballmer's words more carefully, the prediction becomes a lot more ambiguous. What, for example, qualifies as "significant?" What market was he talking about—North American or the world? Was he talking about the market for smartphones or mobile phones in general?

In order to judge forecasts, we can't make assumptions about what the forecast means. There can't be any ambiguity about whether a forecast is accurate or not, and Ballmer's forecast has a lot of ambiguity.

For another example of a forecast that is too ambiguous to be judged right or wrong, Tetlock cites an open letter sent to Ben Bernanke, then-chairman of the Federal Reserve, in November 2010. The letter, signed by several prominent economists and commentators, calls on the Federal Reserve to stop its policy of large-scale asset purchases known as "quantitative easing" because it "risk[s] currency debasement and inflation."

The advice was ignored and quantitative easing continued, but in the years that followed, the U.S. dollar wasn't debased and inflation didn't rise. In 2013, one investor and commentator wrote that the signatories had been proven "terribly wrong." But as Tetlock notes, the letter writers said nothing about the time frame, and so even though currency debasement and inflation hasn't happened yet, it may still happen in the future.

"A Holocaust … will occur"

In 1984, Tetlock was invited to be on a panel of experts convened by the National Research Council, the research arm of the U.S. National Academy of Scientists. The panel included three Nobel laureates and an array of other luminaries. Tetlock, a 30-year-old political psychologist newly promoted to associate professor at the University of California, Berkeley, was invited because of a relevant research project he was working on at the time.

The mission of the panel was to prevent nuclear war. Former Soviet Union leader Leonid Brezhnev had died in 1982 and was eventually replaced by Konstantin Chernenko, who was expected to die soon. The group of experts were all deeply confident that they knew what was happening and where the Soviet Union was heading.

All of the experts believed that the next leader was likely to be another stern Communist Party man, but the liberals and conservatives disagreed about what would happen next. The liberals thought that U.S. President Ronald Reagan's hard-line stance would be met with an equally hard line, and relations between the superpowers would worsen. Conservatives thought the new boss would be the same as the old boss, and the Soviet Union would continue to threaten world peace by invading its neighbors. Both groups were confident in their views.

Then Chernenko died and the Politburo anointed Mikhail Gorbachev as the next general secretary of the Communist Party of the Soviet Union, and he quickly changed direction and liberalized the Soviet Union both politically and economically.

Although few experts on the panel saw this coming, "it wasn't long before most of those who didn't see it coming grew convinced they knew exactly why it had happened, and what was coming next," Tetlock writes. "No matter what had happened the experts would have been just as adept at downplaying their predictive failures and sketching an arc of history that made it appear they saw it coming all along."

Judging judgments

Tetlock says his experience on the panel got him thinking about expert forecasts and whether it was possible to test the forecasting accuracy of political experts.

The more he pondered the challenge, the more he realized why no one had ever seriously tested the accuracy of forecasters. For one thing, there was the problem of timelines. It's much more difficult to judge the accuracy of a forecast when it doesn't include a clear timeline. Another obstacle to judging forecasts is that predictions often rely on implicit understandings of key terms rather than clear definitions. An example would be "significant market share" in Ballmer's forecast.

To illustrate a third obstacle to judging forecasts, Tetlock tells the story of Sherman Kent. "By the time Kent retired from the CIA in 1967, he had profoundly shaped how the American intelligence community does what it calls intelligence analysis—the methodical examination of the information collected by spies and surveillance to figure out what it means, and what will happen next."

For many years, Kent and his colleagues at the Office of National Estimates drew on all the information available to the CIA to make forecasts that might help the U.S. government decide what to do next. In the 1940s, there were fears that the Soviet Union would invade Yugoslavia, and Kent and his colleagues issued this report: "Although it is impossible to determine which course of action the Kremlin is likely to adopt, we believe that the extent of [Eastern European] military and propaganda preparations indicate that an attack on Yugoslavia in 1951 should be considered a serious possibility."

A few days later, Kent was chatting with a senior State Department official who casually asked him, "What did you mean by the

expression 'serious possibility'? What kind of odds did you have in mind?" Kent replied that he felt the odds were about 65 to 35 in favor of an attack. The official was startled, because he and his colleagues had taken "serious possibility" to mean much lower odds.

Then Kent went back to his team and asked each member what he thought the odds were for an invasion. The answers he got varied widely. One analyst gave the odds as 80 to 20, meaning four times more likely than not that there would be an invasion. Another analyst thought the odds were exactly the opposite. The other answers were scattered in between. Kent was floored. A phrase that he thought was so informative turned out to be so vague that it was almost useless.

Kent suggested a solution. To avoid confusion, he suggested that the terms they use have designated numerical definitions, which Kent set out in a chart:

Term: Certain
Meaning: 100 percent certainty

Term: Almost certain
Meaning: 93 percent certainty (give or take about 6 percent)

Term: Probable
Meaning: 75 percent certainty (give or take about 12 percent)

Term: Chances about even
Meaning: 50 percent certainty (give or take about 10 percent)

Term: Probably not
Meaning: 30 percent certainty (give or take about 10 percent)

Term: Almost certainly not
Meaning: 7 percent certainty (give or take about 5 percent)

Term: Impossible
Meaning: 0 percent certainty

Although Kent's idea was simple, it was never adopted. Some people said it felt unnatural or awkward to assign numbers to forecasts. Others said it was tacky to talk explicitly about numerical odds, and it made them sound like bookies. "I'd rather be a bookie than a goddamn poet," was Kent's famous response.

Another objection is that using numbers implies to the reader that the forecast is an objective fact rather than a subjective judgment. But Tetlock says the answer isn't to eliminate numbers. "It's to inform readers that numbers, just like words, only express estimates—opinions—and nothing more."

A more fundamental obstacle to using numbers relates to what Tetlock calls the "wrong-side-of-maybe fallacy." If a meteorologist says there's a 70 percent chance of rain and it doesn't rain, that doesn't mean she's wrong. After all, what she meant was that there was also a 30 percent chance that it wouldn't rain. With just that one day as evidence, it's not possible to judge whether the meteorologist's forecast is correct. "The only way to know for sure would be to rerun that day hundreds of times. If it rained in 70% of those reruns, and didn't rain in 30%, she would be bang on."

This is called calibration. But calibration isn't the whole story when it comes to scoring forecasts, says Tetlock. The other factor to consider is resolution, which is a measure of how often something you say will happen actually happens and vice versa.

Calibration and resolution capture two distinct facets of good judgment. "When we combine calibration and resolution, we get a scoring system that fully captures our sense of what good forecasters should do," he writes.

The math behind this scoring system was developed by a man named Glenn W. Brier in 1950, and so the results are called Brier scores. The Brier score measures the distance between what you forecast and what actually happens. The lower the score is, the better. A perfect score is 0. A hedged 50-50 call, or random guessing in the aggregate, results in a score of 0.5. And the worst possible score, where someone is wrong to the greatest possible extent, is 2.0.

The meaning of the math

In order to judge the accuracy of forecasts to figure out what works and what doesn't, it's not enough to have a reliable scoring system. To interpret the meaning of the Brier scores, we also need two other things: benchmarks and comparability.

What a Brier score means depends on what's being forecasted. Some forecasts are a lot easier to make than others. For instance, if you're forecasting the weather in Phoenix in June, a month when it's typically almost always hot and sunny, you can easily end up with a Brier score close to 0 simply by following a mindless rule like "always assign 100 percent to hot and sunny."

"Hence the right test of skill would be whether a forecaster can do better than mindlessly predicting no change," writes Tetlock.

Another key benchmark is other forecasters, and that requires a level playing field. For instance, forecasting the weather in Phoenix is easier than forecasting the weather in Springfield, Missouri, so it would be unfair to compare the Brier scores of meteorologists in those two cities.

Expert political judgment

This part of the chapter is where Tetlock discusses his landmark study, referred to in this book as "EPJ," which led to his earlier book, *Expert Political Judgment: How Good Is It? How Can We Know?*

EPJ began in the mid 1980s after Tetlock recruited 284 "serious professionals, card-carrying experts whose livelihoods involved analyzing political and economic trends and events." He set out to find the answers to the following questions: How good are the forecasters? Who are the best? What sets them apart?

The forecast questions covered time frames ranging from one to five to 10 years out, and covered a wide range of political and economic topics, some domestic and some international. In some cases, the experts were answering questions in their specific fields of expertise. In others, they were acting as smart, well-informed laypeople. In total, the experts made roughly 28,000 predictions.

And the results …
When the final EPJ results appeared in 2005, the finding that got all the attention was that "the average expert was roughly as accurate as a dart-throwing chimpanzee."

But the other notable finding is that within the results, there were two statistically distinguishable groups of experts. One group failed to do better than random guessing and their longer-range forecasts were actually worse than the chimpanzee. The other group, however, did beat the chimpanzee, although not by a very wide margin. Still, Tetlock says, "however modest their foresight was, they had some."

What Tetlock discovered was that the reason one group did better than the other didn't have anything to do with whether they had Ph.D.s, or whether they had access to classified information. The critical factor wasn't *what* they thought; it was *how* they thought.

Tetlock calls one group the "big ideas" group, and the other "the more eclectic experts" group:

The big ideas group:

- Tended to organize their thinking around big ideas, although they didn't agree on which big ideas were true or false.
- Was united by the fact that their thinking was so ideological. They sought to squeeze complex problems into the preferred cause-effect template and treated what did not fit as irrelevant distractions.
- Was allergic to wishy-washy answers, and kept pushing their analysis to the limit, using terms like "furthermore" and "moreover."
- Was unusually confident and likelier to declare things "impossible" or "certain."
- Was reluctant to change their minds, even when their predictions clearly failed.

The more eclectic experts group:

- Drew on many analytical tools, with the choice of tool hinging on the particular problem they faced.
- Gathered as much information from as many sources as they could.
- Often shifted mental gears when thinking, sprinkling their speech with transition markers such as "however," "but," "although" and "on the other hand."
- Talked about possibilities and probabilities, not certainties.
- More readily admitted when they were wrong and changed their minds.

Tetlock dubbed the big ideas group "hedgehogs" and the eclectic experts "foxes," using terms he found in an essay written by the philosopher Isaiah Berlin, which compared the styles of thinking of great authors through the ages. In the essay, Berlin cites a line from a poem by the Greek warrior-poet Archilochus: "The fox knows many things but the hedgehog knows one big thing."

In the EPJ study, foxes beat hedgehogs. "And the foxes didn't just win by acting like chickens, playing it safe with 60% and 70% forecasts where hedgehogs boldly went with 90% and 100%," Tetlock writes. "Foxes beat hedgehogs on both calibration and resolution. Foxes had real foresight. Hedgehogs didn't."

The EPJ data also revealed an inverse correlation between fame and accuracy. The more famous the expert, the less accurate he was. Tetlock says this isn't because editors, producers and the public go looking for bad forecasters. Instead, it's because hedgehogs tell "tight, simple, clear stories that grab and hold audiences. As anyone who has done media training knows, the first rule is 'keep it simple, stupid.'"

Tetlock goes on to say that audiences tend to find uncertainty disturbing. "The simplicity and confidence of the hedgehog impairs foresight, but it calms nerves."

Dragonfly eye

One of the reasons that the foxes did better than the hedgehogs, Tetlock says, is because they were aggregators of many perspectives, and aggregation is essential for good forecasting.

"How well aggregation works depends on what you are aggregating," Tetlock writes. Aggregating the judgments of many people who know nothing produces a lot of nothing, while aggregating the judgments of people who know a little is better. "But aggregating the judgments of an equal number of people who know lots about lots of different things is most effective because the collective pool of information becomes much bigger."

The best metaphor for this process is the vision of the dragonfly. Each eye is an enormous bulging sphere that is covered with tiny lenses. Depending on the species, there may be as many as 30,000 of these lenses on a single eye, and each lens occupies a slightly different physical space, giving it a unique perspective.

A big reason why foxes are better at forecasting than hedgehogs is because they aggregate perspectives. "Stepping outside ourselves and really getting a different view of reality is a struggle," writes Tetlock. "But foxes are likelier to give it a try. Whether by virtue of temperament or habit or conscious effort, they tend to engage in the hard work of consulting other perspectives."

chapter four:

superforecasters

In this chapter, Tetlock tells the story of how the Intelligence Advanced Research Projects Activity created its unprecedented forecasting tournament, and how it led to the discovery of superforecasters.

The chapter begins with the story of the worst intelligence failure in modern history: the October 2002 report that "Iraq has continued its weapons of mass destruction (WMD) programs in defiance of UN resolutions and restrictions. Baghdad has chemical and biological weapons as well as missiles with ranges in excess of UN restrictions; if left unchecked, it probably will have a nuclear weapon during this decade."

This was soon found to be false. After invading Iraq in 2003, the U.S. searched the country and found nothing. The intelligence community (IC) was humiliated.

As a result, in 2006, IARPA was created with the mission to fund cutting-edge research with the potential to make the intelligence community smarter and more effective.

Two years later, the Office of the Director of National Intelligence asked the National Research Council to form a committee to synthesize research on good judgment and help the IC put that research to good use. Tetlock was asked to sit on the committee because of his EPJ study. The report that the committee later delivered recommended that the IC "rigorously test current and proposed [analytical] methods under conditions that are as realistic as possible. Such an evidence-based approach to analysis will

promote the continuous learning needed to keep the IC smarter and more agile than the nation's adversaries."

In 2010, two representatives from IARPA visited Berkeley to meet with Tetlock and his research partner Barbara Mellers. Tetlock had confidently predicted that the committee's recommendation would gather dust, so he was surprised to learn that IARPA intended to sponsor a major tournament to see who could invent the best methods for making the kinds of forecasts that intelligence analysts make every day.

Examples of possible questions included, "Will an outbreak of H5N1 in China kill more than ten in the next six months?" and "Will the euro fall below $1.20 in the next twelve months?"

Tetlock writes that IARPA was looking for questions in the Goldilocks zone of difficulty: not so easy that any attentive reader of the *New York Times* would know the answer, nor so hard that no one on the planet could.

One big difference between the questions in the proposed tournament and the questions that Tetlock had asked in his study involved time frames. For example, the longest questions in the tournament would be shorter than the shortest questions in Tetlock's original study.

The research teams would be competing against each other as well as an independent control group. The teams had to beat the combined forecast of the control group by 20 percent in the first year and by 50 percent by the fourth year.

In addition, within each team, researchers were encouraged to run experiments to see if certain kinds of training methods helped improve forecasters' accuracy.

Tetlock and his colleagues needed a lot of volunteers, and they put the word out through blogs and professional networks. The first year, several thousand volunteers signed up and about 3,200 passed

through a series of initial tests and started forecasting. Tetlock and his research colleagues called their program the Good Judgment Project.

Tetlock makes the point that it was "bureaucratically gutsy" for IARPA to sponsor this tournament—not because of the cost of the project, which was relatively small considering that the IC's annual budget is around $50 billion—but because of what the results of the tournament might reveal. "Think how shocking it would be to the intelligence professionals who have spent their lives forecasting geopolitical events—to be beaten by a few hundred ordinary people and some simple algorithms."

And yet, that is exactly what happened. The method that the Good Judgment Project used to win IARPA's tournament involved first getting a lot of people to make forecasts and then identifying the 40 or so people who had the best record at making predictions. From there, they asked the entire group to make a lot more forecasts, but gave a little extra weight to those forecasts that were made by the top 40 forecasters.

Resisting gravity but for how long?

Tetlock concludes this chapter by exploring the possibility that it's both skill and luck that explain why certain experts did so well at forecasting. To do so, he introduces the statistical concept of "reversion to the mean," which he says is an indispensable tool for testing the role of luck in performance.

When an activity is dominated by skill, there is slow reversion to the mean. With activities that are dominated by luck (such as a coin toss, for example), there is rapid reversion to the mean.

Tetlock writes: "To illustrate, imagine two people in the IARPA tournament, Frank and Nancy. In year one, Frank does horribly but Nancy is outstanding. … If their results were caused entirely by luck—like coin flipping—then in year two we would expect Frank and Nancy to regress all the way back to 50 percent. If their results were equal parts luck and skill, we would expect halfway reversion: Frank should rise to about 25 percent (between 1 percent and 50

percent) and Nancy [should] fall to around 75 percent (between 50 percent and 99 percent). If their results were entirely decided by skill, there would be no reversion: Frank would be just as awful in year two and Nancy would be just as spectacular."

How did the superforecasters hold up across the years of the tournament? "Phenomenally well," Tetlock says. In years two and three, he and his colleagues saw the opposite of reversion to the mean. The superforecasters actually increased their lead over all the other forecasters.

Tetlock explains that one reason this happened may be because after the first year, the superforecasters were identified and put on teams with their fellow superforecasters. "Instead of reverting towards the mean, their scores got even better. This suggests that being recognized as 'super' and put on teams of intellectually stimulating colleagues improved their performance enough to erase the reversion to the mean we would otherwise have seen."

The conclusion to be made is that the superforecasters weren't just lucky. For the most part, their results reflected genuine skill.

chapter five:

supersmart?

Chapter five is the first of several chapters devoted to what it was that made the superforecasters so good at their predictions: intelligence.

By testing all of the volunteers who signed up for the Good Judgment Project, Tetlock and his colleagues knew that these were intelligent and well-informed people. Regular forecasters scored higher on intelligence and knowledge tests than about 70 percent of the population. The superforecasters placed higher than about 80 percent of the population.

But having the requisite intelligence and knowledge is not enough, Tetlock writes. After all, history is filled with examples of brilliant minds who made forecasts that turned out to be incorrect. "Ultimately, it's not the crunching power that counts," Tetlock writes. "It's how you use it."

Fermi–ize

How many piano tuners are there in Chicago? The Italian American physicist Enrico Fermi gave his students this brainteaser decades before the Internet had been invented. To answer the question, Fermi told his students that they first had to answer the question, "What information would allow me to answer the question?"

Tetlock writes that the question could be answered if you knew these four facts:

- The number of pianos in Chicago
- How often pianos are tuned each year
- How long it takes to tune a piano
- How many hours a year the average piano tuner works

"With the first three facts, I can figure out the total amount of piano-tuning work in Chicago," Tetlock writes. "Then I can divide it by the last and, just like that, I'll have a pretty good sense of how many piano tuners there are in Chicago."

What Fermi understood is that by breaking down the question into a series of questions, you can better separate out the knowable from the unknowable. So although guessing isn't completely eliminated, you can make a more accurate estimate than just randomly thinking of a number.

Tetlock then goes on to further break down each of the four facts into several other questions. For instance, in order to estimate the number of pianos in Chicago, you need to first find out how many people there are in Chicago and then estimate how many people in Chicago own a piano, and so on. This step-by-step process, which Tetlock calls "Fermi-izing," is just what superforecasters do when they are making their predictions.

A murder mystery

In 2012, Yasser Arafat's widow gave permission for her husband's body to be exhumed and tested by agencies in both Switzerland and France for the presence of elevated levels of polonium, a radioactive element that can be deadly if ingested.

Ever since his death in 2004, there had been speculation that Arafat was poisoned. That speculation rose after high levels of polonium were found on some of Arafat's belongings in 2012.

So IARPA asked tournament forecasters, "Will either the French or Swiss inquiries find elevated levels of polonium in the remains of Yasser Arafat's body?" Bill Flack, who knew very little about Middle Eastern politics, tackled the question by first asking, "What would it take for the answer to be yes? What would it take for the answer to be no?"

Flack realized that one of the first questions he needed to answer was whether it was possible to detect polonium on the remains of a man who had been dead for years. Only after determining that this

was possible did Flack move on to the next stage of the analysis, which was to figure out how many different ways Arafat's body could have been contaminated: the more possible ways, the higher the probability. "This was only the beginning, but thanks to Bill's Fermi-style analysis he had already avoided the bait-and-switch tiger trap and laid out a road map for subsequent analysis," Tetlock says. "He was off to a terrific start."

Outside first

When making forecasts, one strategy that forecasters routinely use is to first take the "outside view," a term coined by Daniel Kahneman, a political psychologist and Tetlock's former colleague at UC Berkeley. To illustrate what the outside view means, Tetlock poses a question about whether a particular family has a pet.

To answer that question, many people would zero in on the details of the family and what factors might make it more likely that they'd have a pet. But Tetlock says superforecasters don't start with that approach. Instead, they start with a much broader question: What percentage of American households own a pet?

That's the outside view, or what statisticians call the base rate. In the case of the Arafat-polonium question, the outside view isn't as obvious, Tetlock says, but if you think about the question "Fermi-style," you can come to figure out the outside view.

The inside view

Once you've Fermi-ized the question and consulted the outside view, it's time to consider the inside view, but that doesn't mean you need to slog through endless amounts of information.

"When Bill Flack Fermi-ized the Arafat-polonium answer, he realized there were several pathways to a yes answer," writes Tetlock. So he started with all of the hypotheses, and then broke down each one into the various elements that would need to be true in order for that particular hypothesis to be true. "It's methodical, slow and demanding," writes Tetlock. "But it works far better than wandering aimlessly in a forest of information."

Thesis, antithesis, synthesis

Once superforecasters figure out both the outside view and the inside view, the next step is to merge the two different perspectives into a single vision. But that's not the end, Tetlock says. Superforecasters are constantly looking for other views that they can synthesize into their own.

When a superforecaster explains their thinking to teammates, they're asking them to critique it, which Tetlock says is a very smart move. It's a way to step back from the judgment, scrutinize it from a distance and ask yourself, "Would I be convinced by this if I were somebody else?"

Tetlock says researchers have found that merely asking people to assume that their initial judgment is wrong, to consider why that might be, and to make another judgment produces a second estimate that, when combined with the first, improves accuracy almost as much as getting a second estimate from another person.

Another way to get another perspective is to tweak the wording of the question. For instance, instead of just asking the question, "Will the South African government grant the Dalai Lama a visa within six months?" you can also ask, "Will the South African government deny the Dalai Lama for six months?"

Dragonfly forecasting

Superforecasters are very good at considering several perspectives at the same time, says Tetlock. He calls this the "dragonfly eye" in operation. "Superforecasters pursue point-counterpoint discussions routinely, and they keep at them long past the point where most people would succumb to migraines," he writes.

Superforecasters are "actively open-minded," Tetlock says. "For superforecasters, beliefs are hypotheses to be tested, not treasures to be guarded. It would be facile to reduce superforecasting to a bumper-sticker slogan, but if I had to, that would be it."

chapter six:

superquants?

In chapter six, Tetlock delves into all of the ways in which superforecasters are extremely good with numbers. He starts off by explaining that most superforecasters scored extremely high on a brief test of basic numeracy that asked questions such as, "The chance of getting a viral infection is 0.0005 percent. Out of 10,000 people, about how many of them are expected to get infected?" (The answer is five.)

Many superforecasters have a background in math, science or computer programming. "I have yet to find a superforecaster who isn't comfortable with numbers and most are more than capable of putting them to practical use," Tetlock writes.

But he goes on to reassure readers that you don't have to be a math wizard to be good at forecasting. "While superforecasters do occasionally deploy their own explicit math models, or consult other people's, that's rare. The great majority of their forecasts are simply the product of careful thought and nuanced judgment."

Where's Osama?

Tetlock recounts a scene in the movie *Zero Dark Thirty* that depicts CIA Director Leon Panetta asking his staffers how confident each one of them is that the compound in Pakistan that they've discovered is indeed where Osama bin Laden is hiding. He gets different answers from different people. Some say 60 percent, while others say 80 percent. "This is a clusterfuck, isn't it," the fictional Panetta says, clearly annoyed that not everyone is in agreement.

But the fictional Panetta is wrong to be annoyed, Tetlock says. "An array of judgments is welcome proof that the people around the table are actually thinking for themselves and offering their unique perspectives." He goes on to say that the fictional Panetta should have been delighted that he heard different judgments and should have synthesized the perspectives shared, perhaps giving more weight to judgments from those he respected the most.

Tetlock goes on to say that he asked the real Panetta about that scene in the movie, and Panetta confirmed that something like that did happen. Panetta said judgments varied from people who thought the chances were as low as 30 to 40 percent to others who thought it was 90 percent and above. But rather than being annoyed by the diversity of opinions, Panetta said he welcomed it.

In *Zero Dark Thirty*, a character named Maya emerges as the hero because she boldly asserts, "100 percent, he's there. OK, fine, 95 percent because I know certainty freaks you guys out. But it's 100!" Maya is, of course, right. Bin Laden is there. But Tetlock makes the point that Maya's 100 percent claim was more extreme than the evidence could support, since there was still a small chance that the man in the compound was not Bin Laden.

Probability for the Stone Age

Tetlock makes the point that many people have a hard time grasping the concept of probability. For example, many people think that a 70 percent chance of rain in Los Angeles means that it will rain 70 percent of the day but not the other 30 percent, or that it will rain in 70 percent of Los Angeles but not the other 30 percent.

To grasp the meaning of "a 70 percent chance of rain tomorrow," you need to understand that rain may or may not happen, and if we were able to repeat that same day 100 times, it would rain on 70 of those days and not rain on the other 30.

Former U.S. Treasury Secretary Robert Rubin told Tetlock that his then-deputy Larry Summers would often be frustrated when top policymakers in the White House and Congress would treat an 80

percent probability that something would happen as a certainty that it would. Only when the probabilities were closer to even, say 60-40, would people get the idea, Rubin said.

Probability for the Information Age

Scientists look at probability in a very different way, Tetlock says, because they recognize that nothing is 100 percent certain. All scientific fact is tentative, and nothing is chiseled in stone.

Yet in practice, scientists use the language of certainty because "it is cumbersome whenever you assert a fact to say, 'although we have a substantial body of evidence to support this conclusion, and we hold it with a high degree of confidence, it remains possible, albeit extremely improbable, that new evidence or arguments may compel us to revise our view of this matter.'"

This uncertainty is one reason why scientists prefer to use numbers, Tetlock says. "And those numbers should be as finely subdivided as forecasters can manage," he writes. "The finer-grained the better, as long as the granularity captures real distinctions—meaning that outcomes you say have an 11 percent chance of happening really do occur one percent less often than 12 percent and one percent more than 10 percent outcomes. This complex mental dial is the basis of probabilistic thinking."

Probabilistic thinking is fundamentally different from the three-setting mental dials (yes, no, maybe) that come more naturally to many of us. "Each rests on different assumptions about reality and how to cope with it," writes Tetlock. "And each can seem surpassingly strange to someone accustomed to thinking the other way."

Uncertain supers

Philosophers make a distinction between two different kinds of uncertainty—epistemic and aleatory. Epistemic uncertainty is something that you don't know but that is, at least in theory, knowable. Aleatory uncertainty, on the other hand, is something that you not only don't know but is unknowable. "No matter how much you want to know whether it will rain in Philadelphia one year

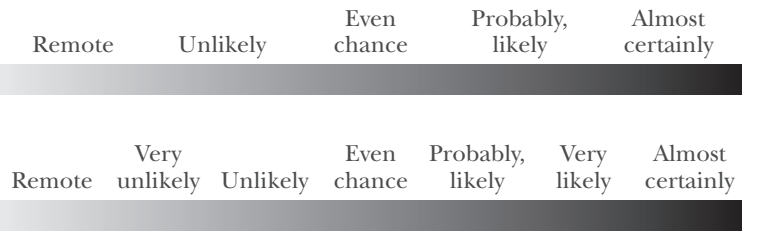
from now, no matter how many great meteorologists you consult, you can't outguess the seasonal averages," Tetlock writes.

Superforecasters grasp this distinction better than most, Tetlock says. When they are given a question that's loaded with irreducible uncertainty, they have learned to be cautious and keep their initial estimates in the "shades-of-maybe zone," 15 percentage points below and above 50 percent. "They know the 'cloudier' the outlook, the harder it is to beat that dart-throwing chimpanzee."

Superforecasters are also more likely than ordinary forecasters to make more granular predictions. For instance, regular forecasters tended to stick to the tens, saying something was 30 percent likely or 40 percent likely. Tetlock says one third of the predictions from superforecasters used the single percentage point scale, meaning they would think carefully and decide that the chance of something happening was 3 percent rather than 4 percent.

But how can we know that the granularity that superforecasters use is meaningful? The proof is in the data, he says. Barbara Mellers found that granularity predicts accuracy. "The average forecaster who sticks with the tens—20 percent, 30 percent, 40 percent—is less accurate than the finer-grained forecaster who uses fives—20 percent, 25 percent, 30 percent—and still less accurate than the even finer-grained forecaster who uses ones—20 percent, 21 percent, 22 percent."

Tetlock says even very sophisticated people and organizations don't push themselves to be as granular as they should be. For instance, the National Intelligence Council asks its analysts to make judgments on a five- or seven-degree scale:



While this degree of granularity is an improvement over the three-setting dial, it still falls short of what the most committed superforecasters can achieve on many questions.

But what does it all mean?
When something unlikely and important happens, it's deeply human to ask why, Tetlock says. Many people believe that everything that happens in our lives happens for a reason, that things that happen are "meant to be"; however, the more scientific view is that everything happens by chance.

To test the theory that superforecasters are much less inclined to see things as fated, Tetlock and his colleagues probed reactions to pro-fate statements such as:

- Events unfold according to God's plan.
- Everything happens for a reason.
- There are no accidents or coincidences.

They also asked the forecasters about these statements:

- Nothing is inevitable.
- Even major events like World War II or 9/11 could have turned out very differently.
- Randomness is often a factor in our personal lives.

They also put the same questions to regular volunteer forecasters, University of Pennsylvania undergraduates and a broad cross section of adult Americans.

"On a 9-point 'fate score,' where 1 is total rejection of it-was-meant-to-happen thinking and 9 is a complete embrace of it, the mean score of adult Americans fell in the middle of the scale. The Penn undergrads were a little lower. The regular forecasters were a little lower still. And the superforecasters got the lowest score of all, firmly on the rejection-of-fate side."

chapter seven:

supernewsjunkies?

Forecasts are based on available information and need to be updated in light of changing information. So it makes sense that superforecasters are people who follow the news closely and update frequently.

One superforecaster named Devyn Duffy is a fantastic updater, Tetlock writes. “Like many superforecasters, Devyn follows developments closely by using Google alerts. If he makes a forecast about Syrian refugees, for example, the first thing he will do is set an alert for ‘Syrian refugees’ and ‘UNHCR,’ which will net any news that mentions both Syrian refugees and the UN agency that monitors their numbers.”

Superforecasters update their forecasts much more frequently, on average, than regular forecasters, which brings up a good point, Tetlock says: Maybe the reason superforecasters are so successful is mainly because they spend so much time watching the news and updating.

But that’s not the whole story, Tetlock says. “For one thing, superforecasters’ initial forecasts were at least 50 percent more accurate than those of regular forecasters. Even if the tournament had asked for only one forecast, and did not permit updating, superforecasters would have won decisively.”

Under- and overreactions

Tetlock says while making frequent updates often improves accuracy, there are two dangers that forecasters can face after they make an initial forecast. One is not giving enough weight to new

information (underreacting) and the other is seeing new information as more meaningful than it is and giving it too much weight (overreacting). Both of these dangers can diminish the accuracy of a forecast.

In some cases, underreaction is simply the result of being so busy that you don't have time to update your forecast after new information becomes available. But in other cases, the reason is more subtle.

To illustrate, Tetlock tells the story of a question that IARPA asked in 2013: Would Japanese Prime Minister Shinzo Abe visit the Yasukuni Shrine? Bill Flack felt strongly that the answer was no. The shrine was founded in 1869 to honor Japan's war dead and conservatives like Abe revere it. But included among the names of the honored dead are about 1,000 war criminals, so visits to Yasukuni by Japanese leaders outrage both the Chinese and Korean governments. The U.S. government constantly urges Japanese prime ministers not to damage relations by visiting.

Based on these facts, Flack was confident that Abe wouldn't go. But then someone close to Abe said, unofficially, that Abe would go. Flack didn't change his forecast because he felt it didn't make any sense for Abe to visit. But on December 26, Abe did visit Yasukuni and Flack's Brier score took a hit.

Tetlock says the mistake that Flack made was ignoring Abe's own feelings. Instead of answering the question, "Will Abe visit?" Flack had answered a different question: "If I were PM of Japan, would I visit Yasukuni?" "Bill recognized that he had unconsciously pulled a bait-and-switch on himself, substituting an easy question for a hard one. Having strayed from the real question, Bill dismissed the new information because it was irrelevant to his replacement question."

Another problem is when a forecaster is reluctant to change their forecast based on new information because that information upsets their beliefs. Consider the comment of General John DeWitt in 1942 concerning the internment of Japanese Americans: "The very

fact that no sabotage has taken place to date is a disturbing and confirming indication that such action will be taken."

But Tetlock writes that superforecasters are less inclined to stubbornly stick to their beliefs in the face of new and conflicting information because they're not experts or professionals, "so they have little ego invested in each forecast."

Superforecasters are also less inclined to overreact to new information and make major changes in their predictions. Tetlock likens their actions to the sailors in Greek mythology: "Scylla was a rock shoal off the coast of Italy. Charybdis was a whirlpool on the coast of Sicily, not far away. Sailors knew they would be doomed if they strayed too far in either direction. Forecasters should feel the same way about under- and overreaction to new information, the Scylla and Charybdis of forecasting. Good updating is all about finding the middle passage."

Captain Minto

In the third season of the IARPA tournament, the top forecaster was a man named Tim Minto, whose Brier score that season was 0.15, which Tetlock compares to being in the league of Ken Jennings' winning *Jeopardy!* 74 games in a row.

Why was Minto so successful? One reason, says Tetlock, is that he updates his forecasts constantly. But the reason he doesn't fall victim to overreacting is because the magnitude of his constant course corrections are very small.

"A few small updates would have put Tim on a heading for underreaction. Many large updates could have tipped him toward overreaction. But with many small updates, Tim slipped safely between Scylla and Charybdis."

Tetlock ends this chapter by emphasizing there is no magical formula for being good at forecasting—just broad principles with lots of caveats. That means superforecasters must understand that the application of these rules requires nuanced judgments.

chapter eight:

perpetual beta

Another defining characteristic of superforecasters is that they have what the psychologist Carol Dweck calls a "growth mindset." What this means is that they believe their abilities are largely the product of effort, and they can grow to the extent that they're willing to work hard and learn.

By contrast, other people have what's known as a "fixed mindset." They believe that they are who they are, and their abilities can only be revealed, not created and developed.

To illustrate the distinction, Tetlock describes a study that Dweck conducted with a group of fifth graders. First, she gave the students relatively easy puzzles to solve, and they all enjoyed working on them. Then she gave the children harder puzzles. Some lost interest and declined an offer to take them home. But others loved the difficult puzzles more than the easy ones. The difference between the two groups, Dweck said, was that the fixed-mindset kids gave up, while the growth-mindset kids knuckled down.

"To be a top-flight forecaster, a growth mindset is essential," according to Tetlock.

Consistently inconsistent

The famous economist John Maynard Keynes was also a very successful investor. Part of the reason he was so successful was that he loved to collect new ideas and was always open to changing his mind.

Tetlock writes: "Legend says that while conferring with Roosevelt at Quebec, Churchill sent Keynes a cable reading, 'Am coming around to your point of view.' His lordship replied, 'Sorry to hear it. Have started to change my mind.'"

For Keynes, failure was an opportunity to identify mistakes, spot new alternatives and try again. When he made bad calls, he didn't retreat from his decisions. Instead, he embraced new ideas. After the crash of 1929, he concluded that stock prices don't always reflect the true value of companies, so an investor should study a company thoroughly. This same approach was being developed by Benjamin Graham, and this "value investing" became the cornerstone of Warren Buffett's fortune.

Trying and failing

Learning to forecast requires trying to forecast, Tetlock says. You can't simply read books about it. It's like learning to ride a bike. What's equally important, he says, is learning from your mistakes. And to do that, you need to get feedback.

Unfortunately, Tetlock says, most forecasters don't get the kind of feedback they need in order to improve. One main reason for this is ambiguous language. For instance, words like "probably" and "likely" make it difficult to judge forecasts.

The other big barrier to feedback, he says, is time. If a forecast spans months or even years, that allows the flaws of memory to creep in. But in addition to ordinary forgetfulness, there is also something called "hindsight bias" or what is sometimes known as the "I knew it all along" effect, when people misremember their forecasts being more accurate than they actually were.

Analyze and adjust

One other common characteristic of the superforecasters is not just their willingness to find out how well they did on a question, but also what they could have done to improve their accuracy. Tetlock says the superforecasters often engage in lengthy online postmortem discussions with their teammates. And individually, too, they spend a

lot of time mulling over what they could have done differently when they were alone.

Often, these postmortems are as careful and critical as the thinking that goes into making the original forecasts. But Tetlock adds that superforecasters tend to be more open to scenarios where they were "almost right" and tended to reject scenarios that suggested they were "almost wrong."

Grit

In addition to having a growth mindset, superforecasters also need to have plenty of grit, which Tetlock defines as "passionate perseverance of long-term goals, even in the face of frustration and failure." The combination of a growth mindset and grit, Tetlock says, "is a potent force for personal progress."

A story about forecaster Anne Kilkenny illustrates this passion: "When the tournament asked a question about refugee numbers in the Central Africa Rebublic, she went to the UN's website and saw the data were a week old. Rather than assume those were the most recent numbers available, she e-mailed the agency and asked how often the data were updated and when the next update could be expected." After that, she got a response, but it was in French, so she wrote back in French asking them to please reply in English. The response she got back was lengthy and in English, and was a great help in making a forecast.

While Kilkenny is not yet a superforecaster, Tetlock adds that given her tenacity, he wouldn't be surprised if she ultimately became one.

Pulling it all together

Tetlock concludes this chapter with a "rough composite portrait of the modal superforecaster."

In their philosophic outlook, Tetlock writes, superforecasters tend to be cautious, humble and nondeterministic (meaning what happens is not meant to be and does not have to happen).

In their abilities and thinking styles, they tend to be actively open-minded, intelligent and knowledgeable, reflective and comfortable with numbers.

In their methods of forecasting, they tend to be pragmatic, analytical, dragonfly-eyed, probabilistic, thoughtful updaters and good intuitive psychologists.

In their work ethic, they tend to have a growth mindset and grit.

chapter nine:

superteams

Tetlock devotes chapter nine to discussing another element that is crucial to the success of the superforecasters: other people.

He begins the chapter by discussing the Bay of Pigs fiasco and the Cuban Missile Crisis—the first an example of when teams can cause terrible mistakes, and the second an example of when a team can sharpen its judgment and accomplish something together that cannot be done alone.

To team or not to team?
When the IARPA tournament was in the planning stages, Tetlock and his colleagues debated whether putting forecasters on teams would improve accuracy. There were strong arguments on both sides of the debate.

Some believed that teams might foster cognitive loafing, allowing some individuals to do little work and benefit from the heavy lifting of others. Another possible problem might be that "groupthink" would occur, preventing the team from questioning assumptions or confronting uncomfortable facts.

On the plus side, people in groups share information and perspectives that can improve their collective forecasts. When forecasters keep questioning themselves and their teammates and welcome vigorous debate, the group can become more than the sum of its parts.

Tetlock said the Good Judgment Project chose to build teams into its research for two reasons: "First, in the real world, people seldom make important forecasts without discussing them with others, so

getting a better understanding of forecasting in the real world required a better understanding of forecasting in groups." The second reason was curiosity. They didn't know the answer and they wanted to, so they ran an experiment.

In the first year of the tournament, before a single superforecaster had been identified, they randomly assigned several hundred forecasters to work alone and several hundred others to work together in teams. The team members wouldn't meet face to face, but could interact via online forums, email, Skype or however else they wanted.

Tetlock also gave the forecasters on teams a primer on teamwork based on the insights gleaned from research in group dynamics.

Superteams

At the end of the first year of the tournament, the results showed that teams were 23 percent more accurate than individuals. So when planning for the second year, Tetlock and his colleagues agreed that teams would be an essential part of the research design.

For the second year, the researchers created teams of superforecasters, with 12 people in each group. At first, there was a lot of "dancing around," recalled superforecaster Marty Rosenthal. "People would disagree with someone's assessment, and want to test it, but they were too afraid of giving offense to just come out and say what they were thinking." But gradually, as they grew more comfortable with each other, the dancing around diminished.

The results of the superteams were impressive. "On average, when a forecaster did well enough in year one to become a superforecaster, and was on a superforecaster team in year two, that person became 50 percent more accurate," Tetlock writes. An analysis in year three got the same results.

Why did the superteams perform so well? Tetlock says it was "by avoiding the extremes of groupthink and Internet flame wars. And by fostering minicultures that encouraged people to challenge each

other respectfully, admit ignorance and request help."

Another feature of the winning teams is that they fostered a culture of sharing. Tetlock's Wharton colleague Adam Grant categorizes people as "givers," "matchers" and "takers." Givers contribute more to others than they receive in return, matchers give as much as they get, and takers give less than they take. Grant's research shows that when someone is a giver, it improves the behavior of others, including the giver, which explains why givers tend to come out on top.

Finally, the diversity of the team's perspectives is also important. "Combining uniform perspectives only produces more of the same, while slight variations will produce slight improvements. It is the diversity of the perspectives that makes the magic work."

chapter ten:

the leader's dilemma

In this chapter, Tetlock sets out to resolve the apparent contradiction between the demands of good judgment and effective leadership.

If you ask people to list the qualities of an effective leader, you'll find near universal agreement on these three basic points:

- Leaders must be reasonably confident and instill confidence in those they lead.
- Leaders must be decisive. They can't ruminate endlessly. They need to size up a situation, make a decision and move on.
- Leaders must deliver a vision—a goal that everyone strives together to achieve.

Now look at the style of thinking that produces superforecasting, and consider how it squares with what leaders must deliver, Tetlock writes. "How can leaders be confident, and inspire confidence, if they see nothing as certain?" How can they be decisive if their thinking is slow, complex and self-critical? How can they act with relentless determination if they readily adjust their thinking in light of new information?

Tetlock admits that this appears to be a serious contradiction. Leaders must be good forecasters, but it seems that what is required to succeed at one role might undermine the other. But he says the contradiction between being a superforecaster and a superleader is more apparent than real, and to support that argument, he goes on to give examples of approaches to leadership from history that are very much in keeping with the skills and traits of superforecasters:

Helmuth von Moltke: This 19th-century Prussian general once wrote, "In war, everything is uncertain." Moltke's writings on war profoundly shaped the German military that fought the two world wars, but Moltke was no Napoleon, says Tetlock. "He never saw himself as a visionary leader directing his army like chess pieces. His approach to leadership and organization was entirely different."

Moltke believed that you could never entirely trust your plan. He wrote that it was impossible to lay down binding rules that apply in all circumstances, and that improvisation was essential because two cases will never be exactly the same.

Dwight D. Eisenhower: "Like Moltke, Eisenhower knew nothing was certain," writes Tetlock. The first thing the 34th U.S. president did after giving the irrevocable order to go ahead with the D-Day invasion was to write a note taking personal responsibility, which was to be released in the event of failure. Eisenhower also expected his officers to engage in open debate, and he respected well-founded criticisms, Tetlock writes.

David Petraeus: When Tetlock spoke to Petraeus about his philosophy of leadership, it was easy to hear echoes of Moltke, he writes. To develop flexible thinking, the military officer and former director of the CIA pushed people out of their intellectual comfort zones.

Examples from the business world

Tetlock also writes about business leaders who exhibit many of the same qualities that superforecasters do. For instance, he includes a quote from 3M Senior Vice President William Coyne that sounds very much like General Moltke: "We let our people know what we want them to accomplish. But—and it's a very big but—we do not tell them how to achieve those goals."

One of the 14 leadership principles that is drilled into new employees at Amazon is: "Have backbone, disagree and commit." And finally, Tetlock says when Walmart found it was building stores faster than it could develop store managers, the company created a "leadership academy" to get prospects ready for promotion sooner.

"The academy is modeled after military academies with the 'mission command' philosophy at its foundation," he writes.

A peculiar type of humility

Then there's the vexing question of humility. The leaders and experts mentioned throughout the book are self-assured people with very healthy egos.

How does all of that square with the need for a forecaster to be humble? Tetlock says the answer is that the humility that's required for good judgment is not self-doubt, or the sense that you are untalented or unworthy. Rather, it's a trait known as intellectual humility. "It's a recognition that reality is profoundly complex, that seeing things clearly is a constant struggle, when it can be done at all, and that human judgments must therefore be riddled with mistakes."

That's why it's possible to think highly of yourself and still be intellectually humble, Tetlock says. "Intellectual humility compels the careful reflection necessary for good judgment; confidence in one's abilities inspires determined action."

chapter eleven:

are they really so super?

In this chapter, Tetlock presents what he believes are the two strongest challenges to his research.

Scope insensitivity

The first challenge comes from cognitive psychologist Daniel Kahneman, who suggests that even superforecasters are unable to resist a bias known as scope insensitivity.

"Kahneman first documented scope insensitivity 30 years ago when he asked a randomly selected group in Toronto, the capital city of Ontario, how much they would be willing to pay to clean up the lakes in a small region of the province," Tetlock writes. On average, people said they would pay about $10. He then asked a different randomly selected group how much they would be willing to pay to clean up every one of the 250,000 lakes in Ontario. The answer was the same: $10.

In subsequent studies, Kahneman gave different groups different figures for the number of birds who drowned in oil ponds each year. One group was told it was 2,000 birds, a second group was told it was 20,000, and a third group was told the number was 200,000. Yet, in each case, the average amount that people said they'd be willing to pay was around $80. That's an example of scope insensitivity. Instead of answering the question that was asked, they instead answered, "How bad does this make me feel?" And that's why the answer was roughly the same.

To test Kahneman's theory, Tetlock and his colleagues asked a randomly selected group of superforecasters this question: "How

likely is it that the Assad regime will fall in the next three months?" Another group was asked how likely it was in the next six months. Then they did the same experiment with regular forecasters.

Kahneman predicted widespread "scope insensitivity": that the time frame would make no difference to the final answers. The results showed that, as Kahneman had predicted, the vast majority of regular forecasters were scope insensitive. Regular forecasters said there was a 40 percent chance the Syrian president's regime would fall over three months, and a 41 percent chance it would fall over six months. But the superforecasters didn't fall into that same trap. They put the probability of Assad's fall at 15 percent over three months and 24 percent over six months.

Tetlock says he wishes he could take credit for the superforecasters' ability to shake off scope insensitivity, but he suspects that superforecasters are so well practiced at System 2 corrections that these techniques have become almost habitual.

Enter the black swan

Another criticism comes from Nassim Taleb, a former Wall Street trader, who believes that Tetlock's research project is misguided. Taleb introduced the "black swan" concept, which is "an event so far outside experience we can't even imagine it until it happens," Tetlock writes.

What's more, Taleb believes that it is black swans and only black swans that determine the course of history. What this means, Tetlock goes on to say, is that "IARPA has commissioned a fool's errand. What matters can't be forecast and what can be forecast doesn't matter."

In other words, Taleb believes Tetlock's Good Judgment Project is "premised on misconceptions, panders to desperation and fosters foolish complacency."

Tetlock believes Taleb's critique introduces another false dichotomy: "'Forecasting is feasible if you follow my formula' versus 'forecasting is bunk.'" To dispel the dichotomy, Tetlock says it's necessary to

figure out what exactly a black swan is. The strict definition is something that's literally inconceivable before it happens. But Tetlock says if that's the case, many events that are dubbed black swans are actually gray.

For instance, many people point to the 9/11 terrorist attacks as a prototypical black swan. But Tetlock goes on to say that 9/11 was not unimaginable. In fact, he writes, "the danger was so well known in security circles that in August 2001, a government official asked Louise Richardson, a Harvard terrorism expert, why no terrorist group had ever used an airplane as a flying bomb." Richardson later wrote that her answer was that "the tactic was very much under consideration and I suspected that some terrorists groups would use it sooner rather than later."

All that said …

Tetlock concludes the chapter by making the point that he, Taleb and Kahneman all agree that there is no evidence that geopolitical or economic forecasters can predict anything 10 years out beyond obvious statements such as "there will be conflicts." But he adds that if you have to plan for a future beyond the forecasting horizon, you should plan for surprises.

chapter twelve:

what's next?

In this final chapter, Tetlock talks about the future of forecasting.

The kto-kogo status quo

In chapter one, Tetlock says the exclusive goal of forecasting should be accuracy. But here he acknowledges that accuracy is sometimes one of many goals, and sometimes it's irrelevant.

For example, consider Nate Silver's forecasts in the months leading up to the 2012 U.S. presidential election. Silver's forecasts consistently had incumbent President Barack Obama leading Republican nominee Mitt Romney, even when other polls said the race was too close to call. Republicans reviled Silver and accused him of bias. Democrats defended his integrity.

Fast-forward to March of 2014, when Silver issued forecasts that suggested the Republicans would take control of the Senate in that November's midterm elections. Many Democrats had a change of heart, and some even passed around old forecasts of Silver's that had failed. "Same forecaster, same track record—but when his forecasts ceased to align with partisan purposes he was demoted from prophet to stumblebum," Tetlock writes.

The story illustrates how forecasting can be co-opted to advance the interests of a particular party or group. Vladimir Lenin believed that politics was nothing more than a struggle for power, or as he put it, "kto, kogo," which was Lenin's shorthand for "who does what to whom." So it follows that the goal of a forecaster isn't to see what is coming up, but to advance the interest of the forecaster and the forecaster's tribe.

Change

But Tetlock is hopeful about the future. He gives an example from the world of medicine to show how it's possible for people to move beyond a system that overcomes entrenched interests and toward a singular goal of determining whether policies do what they are supposed to do.

A decade ago, a Boston-area doctor named Ernest Codman proposed that hospitals should record what ailments patients had when they came in, how they were treated and the end result of each case. Hospitals initially hated the ideas because they'd have to pay for record-keepers, and doctors didn't like it because they felt that keeping score could only damage their reputations.

But eventually the plan was adopted, and today evidence-based policy is modeled after evidence-based medicine, with the goal of subjecting government policies to rigorous analysis so that legislators will know whether policies do what they are supposed to do. Tetlock says if kto-kogo were the decisive force in human affairs, evidence-based medicine would never have been adopted.

Today, that same evidence-based thinking has been adopted not just by politicians but also by charitable organizations, the professional sports world and other industries.

"Perhaps the best reason to hope for a change is the IARPA tournament itself," Tetlock writes. He says if someone had asked him a decade ago to list the organization that most needed to get serious about forecasting, the intelligence community would have been at the top of the list.

The humanists' objection

Although he is cautiously optimistic that his work may contribute to an evidence-based forecasting movement, Tetlock notes that it's not a prospect that everyone embraces, and he says if he had made a different decision early in his career, he might not be embracing it himself.

As a 22-year-old Canadian who had just graduated from the University of British Columbia, Tetlock was deciding whether to go to the United States and commit to the scientific method or accept a scholarship to study the humanities at Oxford. If he had decided on a career in the humanities, he probably would believe that although numbers are fine and useful things, "we must be careful not to be smitten with them."

Tetlock writes: "'Not everything that counts can be counted,' goes a famous saying, 'and not everything that can be counted counts.'" He goes on to say that he believes there's a lot of truth in that perspective, and that too many people treat numbers like sacred totems.

The question that counts

Asking good questions is a critical dimension of good judgment, Tetlock says, and he defines a good question as "one that gets us thinking about something worth thinking about."

But Tetlock believes that superforecasters aren't always good at asking the good questions, and vice versa. He believes that rather than dwell on each other's weaknesses, people who are good at asking the right questions and people who are good at answering them should acknowledge each other's complementary strengths and work together.

One last idea

Tetlock ends the chapter by discussing another collaboration that he hopes to see in the future, and that involves forecasters with opposing views working together "to identify what they disagree about and which forecasts would meaningfully test those disagreements."

The process is likely to take many questions and many answers, Tetlock writes. And it's possible that the final outcome won't completely resolve the dispute. "A major point of view rarely has zero merit," Tetlock writes, "and if a forecasting contest produces a split decision we will have learned that the reality is more mixed than either side thought. If learning, not gloating, is the goal, that is progress."

When strident voices dominate debates, often neither side is interested in adversarial collaboration, Tetlock notes. But he says we shouldn't make the cynic's mistake of thinking that those who dominate the debate are the only ones we are debating. There are less voluble and more reasonable voices that we should be listening to, he says. If we let these voices design clear tests of their beliefs, we can all watch and see the results "and get a little wiser. All we have to do is get serious about keeping score."

appendix

ten commandments for aspiring superforecasters

Here, Tetlock distills the key themes of the book to give aspiring forecasters a clear set of guidelines:

1. Triage.
2. Break seemingly intractable problems into tractable sub-problems.
3. Strike the right balance between inside and outside views.
4. Strike the right balance between under- and overreacting to evidence.
5. Look for the clashing causal forces at work in each problem.
6. Strive to distinguish as many degrees of doubt as the problem permits but no more.
7. Strike the right balance between under- and overconfidence, between prudence and decisiveness.
8. Look for the errors behind your mistakes but beware of rearview-mirror hindsight biases.
9. Bring out the best in others and let others bring out the best in you.
10. Master the error-balancing bicycle.
11. Don't treat commandments as commandments.

About World 50

World 50 is the premier resource for senior executives from globally respected organizations to privately and candidly share ideas, solutions and collaborative discovery.

The World 50 community provides unrivaled access to and collaboration with more than 900 senior executives from 500 globally respected organizations on six continents. Membership provides unparalleled access to world-class gatherings as well as year-round peer-to-peer and team-to-team collaboration, delivering insights found nowhere else. Intimate participation with remarkable practitioners and expert thinkers creates a candid dialogue on leading and growing significant enterprises in a global economy.