Scientific improvement of decision-making and risk management

Yngve Hoiseth and Martin Holten, Empiricast

February 2020

Executive Summary

Bill Gates, successful entrepreneur and philanthropist, said:

> I have been struck by how important measurement is to improving the human condition. You can achieve incredible progress if you set a clear goal and find a measure that will drive progress toward that goal. . . . This may seem basic, but it is amazing how often it is not done and how hard it is to get right [15, p. 15].

In this white paper, we discuss how organizations can reach their goals through measurement and continuous improvement as effectively and efficiently as possible. It is intended for professionals in business, government and non-profits involved with

- risk management;
- investment decision-making, broadly defined — e.g. choosing a supplier, investing in a company or hiring an employee; and
- strategy development.

We make the following claims.

1. Decision-making and risk management depend critically on forecasting accuracy.
2. Most organizations are unscientific in their approach to forecasting.
3. They can significantly improve their forecasting accuracy by adopting a more scientific approach.
4. It follows that **most organizations can significantly improve their decision-making and risk management by adopting a scientific approach to forecasting**.

If you have any questions or feedback, feel free to reach out to any of us:

- Yngve Høiseth (yngve@empiricast.com/+47 988 18 917)
- Martin Holten (martin@empiricast.com/+47 458 04 565)

Contents

Chapter 1

Background

1.1 The miracle of modern medicine

> When George Washington fell ill in 1799, his esteemed physicians bled him relentlessly, dosed him with mercury to cause diarrhea, induced vomiting, and raised blood-filled blisters by applying hot cups to the old man's skin [15, p. 26].

How could they subject him to such torture? The answer is simple: They thought it would help. But how could they think that? You see, for thousands of years, the medical profession was deeply unscientific. Doctors trusted their experience and intuition, and were way more confident in their abilities than the evidence supported. It was so bad that many patients would probably have been better off had they not seen any doctors at all. Consider Galen, a second-century physician, who infamously wrote:

> All who drink of this treatment recover in a short time, except those whom it does not help, who all die. It is obvious, therefore, that it fails only in incurable cases [15, p. 27].

This sloppy thinking stands in stark contrast to today's medical research, which saves countless lives every day. Progress happened in many small steps, but they all depended on a big one: Making the medical profession scientific. In other words, doctors had to embrace uncertainty — accept that they did not know — and work systematically to improve their methods.

While it may seem obvious today, using scientific methods like randomized controlled trials[1] was controversial well into the twentieth century, as illustrated by the *Lancet* in 1921:

> Is the application of the numerical method to the subject-matter of medicine a trivial and time-wasting ingenuity as some hold, or is it an important stage in the development of our art, as others proclaim? [15, p. 28]

It took the efforts of many pioneers to revolutionize medicine. One of them was Ernest Amory Codman in Boston. He proposed what he called the *End Result System*, in which hospitals would record diagnoses, treatments and results. This would help them improve. And today,

> hospitals do much of what Codman demanded, and more. ... [But at the time,] hospitals hated it. They would have to pay for record keepers. And the physicians in charge saw nothing in it for them. They were already respected. Keeping score could only damage their reputations [15, p. 256].

Thankfully, the likes of Codman eventually won.

To be clear, the medical profession is by no means perfect: As surgeon, writer and public health researcher Atul Gawande documented in his 2011 book *The Checklist Manifesto: How to get things right*, doctors can still be astonishingly slow to pick up on methods proven to save lives [7]. And there are to this day plenty of charlatans out there. But, for all its shortcomings, modern medicine truly is one of humanity's greatest successes.

What took so long? One problem, as mentioned above, is that when individuals have a lot of prestige, they are reluctant to change the system. But there are deeper reasons, too: Humans evolved to survive and reproduce in the wild, not to understand statistics. The groundbreaking work by Daniel Kahneman and Amos Tversky showed us *cognitive biases* — how we make shortcuts to save time and energy. These shortcuts are rational in many cases, but can seriously distort our thinking in others [11]. Galen, the second-century physician mentioned above, would have made more progress had it not been for his *confirmation bias.*

Also, people who are skeptical of quantification sometimes have valid points, as measurement can be unnecessary, misguided or even harmful. For example,

[1] In order to figure out whether an intervention such as a drug or lifestyle change has the intended effect, it's not enough to look at what happens when doctors prescribe them — it's too hard to separate cause and effect. You need to randomly divide patients into two groups (treatment and control), give one the prescription and the other a placebo. Then, you can estimate the effect of the intervention. The larger your sample, the more certain you can be.

when British hospitals needed to cut their emergency room waiting times, their ambulances circled the block until the waiting time was below the required four hours. Managers need to be careful not to let metrics distort incentives [13].

How can professionals in different fields learn from the success of the medical establishment? To find out, let's look at US national intelligence at the beginning of the millennium.

1.2 The weapons that weren't

In October 2002, the US Director of National Intelligence stated:

> We judge 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 [10, p. 5].

For this reason (and maybe others), the US invaded. Their armed forces

> turned Iraq upside down looking for WMDs but found nothing. It was one of the worst — arguably *the* worst — intelligence failure in modern history. ... The bureaucracy was shaken to its foundation [15, p. 82-85].

To help the intelligence community improve, the Intelligence Advanced Research Projects Activity (IARPA) was established in 2006 [1]. A couple of years later,

> the Office of the Director of National Intelligence asked the National Research Council to establish a committee to synthesize and assess evidence from the behavioral and social sciences relevant to analytic methods and their potential application for the U.S. intelligence community [5, p. xiii].

The committee's report, *Intelligence Analysis for Tomorrow: Advances from the Behavioral and Social Sciences*, recommended that the intelligence community

> adopts scientifically validated analytical methods and subjects its methods to performance evaluation. To implement this recommendation, the committee offers three immediate actions:

1. institutionalize an “Analytical Olympics” to test competing analytic methods and foster a culture that values continuous improvement;
2. begin to assess how well-calibrated[2] individual analysts are and provide them with appropriate feedback; and
3. create a research program that reviews current and historic analyses comparing alternative methods under real world conditions [5, p. 3].

1.3 The Analytical Olympics

IARPA took the recommendations seriously. They funded and facilitated a prediction tournament in which external teams were compared to professional intelligence analysts and a control group in order to find ways to improve forecasting accuracy. They competed on questions like:

- Will the president of Tunisia flee to a cushy exile next month?
- Will an outbreak of H5N1 in China kill more than ten in the next six months?
- Will the euro fall below $1.20 in the next twelve months?

Participants would answer with percentages — i.e., “the euro is 42% likely to fall below $1.20 in the next twelve months.” They could update their predictions as new information became available, and they were scored relative to the other participants [15, p. 88].

Professor Philip Tetlock at the University of Pennsylvania has studied prediction accuracy since the 80s. He assembled a team of colleagues and recruited volunteers to help them. He tells the story in his 2015 book *Superforecasting: The Art and Science of Prediction*:

> In year 1, [we] beat the official control group by 60%. In year 2, we beat the control group by 78%. . . . [Our] forecasts even beat those of professional intelligence analysts inside the government who have access to classified information — by margins that remain classified. . . . Of course it would be wonderful to have a direct comparison between superforecasters and intelligence analysts, but such

[2] *Calibration* in this context means that, on average, individuals are neither over- nor underconfident. If you, for example, say that you’re 60% confident of something, it should over time be true 6 out of 10 times.

> a thing would be closely guarded. However, in November 2013, the *Washington Post* editor David Ignatius reported that “a participant in the project” had told him that the superforecasters “performed about 30 percent better than the average for intelligence community analysts who could read intercepts and other secret data” [15, p. 17, 90 and 95].

During the the research program, Tetlock’s team conducted a wide array of experiments in order to figure out how to improve forecasting accuracy. For example, they found that “on average, teams were 23% more accurate than individuals” [15, p. 201]. They also conducted psychometric tests of participants in order to see what traits contributed to high performance[3] and studied the role of luck.[4]

[3] The most important thing to become a great forecaster is commitment to “belief updating and self-improvement.” This commitment is roughly three times as important as its closest rival, intelligence [15, p. 192]. Note, however, that the sample was skewed — for example, the participants had higher-than-average intelligence [15, p. 109].

[4] If you have enough participants in a guessing game, someone will win by chance alone. So to figure out whether it was all luck or not, Tetlock looked at how results correlated between years. If winners were just lucky, there would be no correlation. If there was just skill, we would see perfect correlation. For superforecasters, the correlation was 0.65 on a scale from 0 to 1, showing significant skill [15, p. 104]. If you did the same in your organization, you would probably end up with different numbers, depending on who participates, what they forecast, the resources that are available to them and so on.

Chapter 2

State of the union

What about fields other than national intelligence? Unfortunately, forecasting often fails there, too. Project management researcher Bent Flyvbjerg at Oxford University coined the *iron law of megaprojects*[1]:

> Over budget, over time, over and over again. Nine out of ten such projects have cost overruns. Overruns of up to 50 percent in real terms are common, over 50 percent not uncommon. . . . Overrun is a problem in private as well as public sector projects, and things are not improving; overruns have stayed high and constant for the 70-year period for which comparable data exist. . . . Large-scale [Information and Communication Technology (ICT)] projects are even more risky. One in six such projects become a statistical outlier in terms of cost overrun with an average overrun for outliers of 200 percent in real terms [6, p. 9-10].

Widely considered the most legendary of all planning disasters, the Sydney Opera House was originally supposed to be finished in 1963 and cost $7 million. A scaled-down version finally opened in 1973 for $102 million — ten years late at 15 times the predicted cost [8].

There are funnier examples, such as a forecast Steve Ballmer made in 2007 when he was CEO of Microsoft:

> There's no chance that the iPhone is going to get any significant market share. No chance [15, p. 46].

[1] Megaprojects are large-scale, complex ventures that typically cost a billion dollars or more, take many years to develop and build, involve multiple public and private stakeholders, are transformational, and impact millions of people [6, p. 3].

Low accuracy is a problem because decision-making and risk management depend critically on forecasting. For example, if you are running a for-profit company considering an investment in another company, you are trying to forecast what will happen if you decide to invest:

- What will your financial return on investment be? Will it be more than for alternative investments?
- What will the value of synergy effects be? Will your investment help you protect your position in a relevant market?
- What are your risks other than direct investment losses? Will your investment, for example, cause you a public relations scandal?

Some uncertainty when answering questions like these is unavoidable. (We're not aware of good evidence that anyone can predict complex systems more than five years out.) But some of the uncertainty can be removed, and a lot of research and experience tells us that organizations can get better over time.

In order to improve decision-making and risk management, we need to get serious about improving forecasting. And we can't rely on intuition and experience alone, as they often fail us. In fact, we can't even reliably remember what we thought at the time we made the decision: When Tetlock in an earlier research project asked political experts to recall their estimates on a question asked four years before, they

> recalled a number 31 percentage points higher than the correct figure. One expert thought in 1988 that the Communist Party had a 20% chance of losing their monopoly power in the Soviet Union in the next five years. When asked to recall his estimate, he thought he had said 70% [15, p. 184].

Our memories don't serve us well enough.

Chapter 3

A blueprint for improvement

What does it take to improve? There are two main steps:

1. Measure
2. Learn

3.1 Measurement

Measurement, as defined by statistician Douglas W. Hubbard, is "a quantitatively expressed reduction of uncertainty based on one or more observations" [9, p. 31]. Because measurement is merely about *reducing* uncertainty, we don't need perfection for it to count as measurement.

To illustrate, imagine that you're near a thunderstorm. You don't know how far away it is and whether it's getting closer. But you know that sound travels approximately one-third of a kilometer per second. So you start counting the seconds from flashes to thunders, and multiply the counts by three to get the approximate distance in kilometers to the storm. This is by no means a perfect way to measure the distance, but you have significantly reduced your uncertainty, and you have used numbers to do so. That, by definition, is measurement.

When trying to predict the future, measurement takes a different form. In the example above, we asked about the financial return on your hypothetical investment opportunity. First, you need to figure out exactly what you mean by *financial return*. Your return would be the price you sold for (or could have sold for) minus the price you bought for.

To specify your forecast, you can use a confidence interval: Let's say that you're 90% sure that your return will be between $1 million and $8 million two years from now. That is a measurement of what you think is going to happen.[1]

There are better and worse ways to go about forecast measurement. Often, organizations use needlessly limited measurement methods: They sometimes use words such as "probable" or "unlikely," a traffic light metaphor (green, yellow and red) or a scale, e.g. from 1 to 5 or from 1 to 7. There are two problems with such approaches:

1. People attach different meaning to different words or points on the scale. If something turns out to be true 5 out of 10 times when you said "probable," are you better or worse than someone whose ratio was 7 out of 10? (Figure 3.1 illustrates this problem.)
2. Second, using a finer scale can improve accuracy. In the IARPA tournament, they used whole percentage points: 0%, 1%, 2% and so on, all the way up to 100%. To figure out whether the whole scale really was relevant, Tetlock experimented with rounding, e.g. from 42% to 40%. Their best forecasters "lost accuracy in response to even the smallest-scale rounding, to the nearest [5 percentage points]" [15, p. 145].

When predictions are made using the appropriate amount of resolution, they can be compared to real events. But, more often, "forecasts are made and then . . . nothing. Accuracy is seldom determined after the fact and is almost never done with sufficient regularity and rigor that conclusions can be drawn" [15, p. 14]. To get better, you need to compare your forecasts to what actually happened. In our investment example, that means calculating the return after the two years have gone by.

But what, exactly, does it mean to *compare* predictions and results? In practice, this means *scoring* predictions. The more accurate a prediction, the better its score should be. And, if there are multiple forecasters, scores can be compared against a benchmark (e.g. the median score) in order to account for the fact that some questions are more difficult than others.

In the 1940s, Glenn Brier developed a method to score weather forecasts [4].[2] A *Brier score* measures the distance from the truth. A perfect forecast yields

[1] If you are *calibrated*, the true number will over time end up in your range 9 out of 10 times.

[2] Like medicine, weather forecasting came early to the science party. It's easy to take it for granted, but generations of weather forecasters have worked hard and smart to get as accurate as they are today.

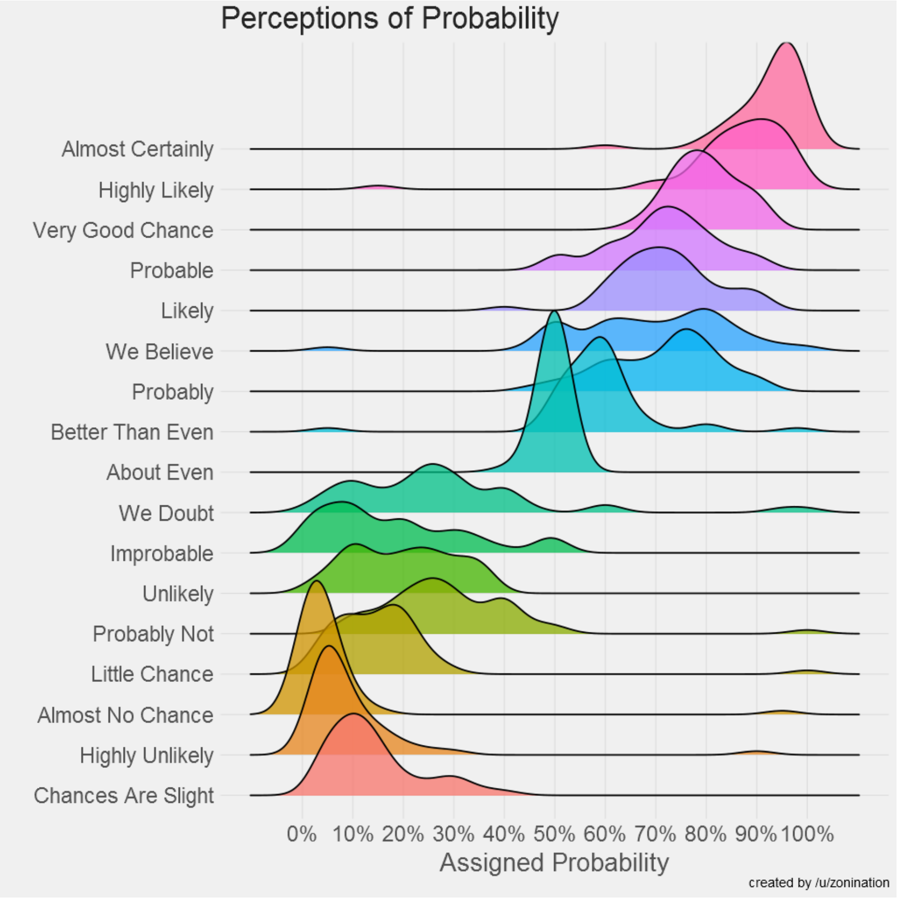

Figure 3.1: Perceptions of Probability [14]

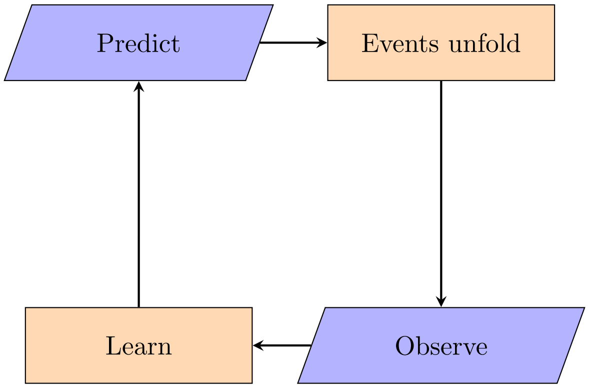

Figure 3.2: Ideal feedback loop

a score of 0 and a complete miss scores 2. If you just give all options the same probability (fifty-fifty if there are two), you get 0.5.[3] Today, Brier's scoring method is the most commonly used across a variety of fields. Scoring confidence intervals is less straight-forward, but good approximations exist.

When these two things — measurement and comparison – are in place, you can start learning.

3.2 Closing the loop

An ideal feedback loop is illustrated in figure 3.2. As we have discussed, *Learn* is usually missing or severely lacking. Even *Observe* is often skipped or done poorly. By scoring predictions, you can close the loop.

The learning can take many forms, such as:

- Give individuals feedback on their performance on individual questions
- Incentivize individuals by facilitating friendly competition
- Share lessons learned and develop best practices
- Build and improve models to aggregate individual predictions [3]

[3] An interesting, almost philosophical, question is what should happen to someone who says they're absolutely certain of something that ends up not happening — e.g. *I'm 100% certain. Hillary Clinton is going to win the 2016 election.* Because Brier scoring limits how bad a score can get — 2 is the worst you can do in the most common implementation — it's possible to recover from such a bad miss. Other methods are less forgiving.

- Experiment with different approaches (like medical researchers experiment with different drugs)
- Automate at least part of the work of making predictions by using software to perform repeatable tasks

3.3 Obstacles

If closing the loop is such low-hanging fruit, why doesn't everyone pick it? Cultivating this kind of scientific management may seem simple in principle, but it is by no means easy. Here are some of the difficulties organizations can encounter.

3.3.1 Questions

In order to get good answers, you need good questions. And asking good questions can be difficult. In the investment opportunity example introduced above, what exactly should you try to forecast? The big question that you're really interested in is something along the lines of "will this investment pay off?" But that's both difficult to predict and difficult to determine the correct answer to.

So you need to make it more concrete and break it down into smaller parts, like we did in our example:

- What will your financial return on investment be? Will it be more than for alternative investments?
- What will the value of synergy effects be? Will your investment help you protect your position in a relevant market?
- What are your risks other than direct investment losses? Will your investment, for example, cause you a public relations scandal?

If you're confident that your sub-questions cover all the possible relevant outcomes, you can aggregate them using *decision tree analysis* [12]. However, in our case, there may be other effects that our questions don't cover. Maybe being associated with the investment strengthens our brand or motivates our staff? Or maybe the company we invest in hires some of our key employees? In such cases, we can use *Bayesian question clustering*, which is more flexible but less straightforward and more subjective [15, p. 263]. The two methods can also be combined for different parts of the tree.

If the things you care about can be expressed with time series, you save a lot of trouble: You don't have to come up with questions and alternatives, and you don't have to manually create and resolve the individual questions. Instead, you need to determine the time series you care about and then ask for forecasts. Optionally, you can also set specific target dates and desired confidence intervals, but that can be left to the respondents.

3.3.2 Systems

Existing systems and processes often stand in the way of better ones. Some examples:

- Risk management systems that lock users into the too coarse-grained and vague traffic light metaphor mentioned above.
- Legacy IT systems that don't speak to each other, making it difficult to get good data on predictions and actual events.
- Difficult-to-use systems that make even the task of gathering predictions daunting.

3.3.3 Incentives

As mentioned above, people often have established prestige which they may lose if their accuracy is measured. That's one of the reasons why media personalities often cloud their forecasts in vague language. If you say that something "may happen," you can't be proven wrong no matter what happens. The same is true in organizations.

A different variant is that people in the organization may deliberately adjust forecasts in order to further goals other than accuracy. It may be a manager that sets an optimistic deadline in order to motivate their subordinates, a broker that jacks up the forecast of an investment in order to get the sale, or a software developer that provides an optimistic deadline in order to seem competent.

Some things will be handled almost by default when measuring forecasts — the notoriously optimistic broker will be exposed — but human factors still need to be kept in mind. And they depend on the organizational context and implementation details. For example, a low-trust environment may call for anonymous forecasts. Also, if accurate predictions are too strongly incentivized and forecasters can affect the outcome, they can become self-fulfilling prophecies. You wouldn't want an employee cutting corners or working slowly just to hit their time estimate.

Chapter 4

Putting it into practice

In addition to the overarching goal of increased accuracy, we believe that organizations that decide to work systematically to improve their forecasting accuracy will benefit in other ways as well. Some examples:

- Quicker response to new information
- Reduced meeting activity
- More rapid employee development and increased motivation
- Strengthened culture for continuous improvement
- Reduced loss of knowledge when employees quit

In practice, well-designed software is necessary when working collaboratively to improve forecasting accuracy.

Chapter 5

Conclusion

Above, we have discussed a general method for improving organizational management. In essence, it's simple: "Forecast, measure, revise. Repeat" [15, p. 14]. In an era of accelerating change, it's more important than ever for organizations to stay on top of their game [2]. We conclude that:

1. Decision-making and risk management depend critically on forecasting accuracy.
2. Most organizations are unscientific in their approach to forecasting.
3. They can significantly improve their forecasting accuracy by adopting a more scientific approach.
4. It follows that **most organizations can significantly improve their decision-making and risk management by adopting a scientific approach to forecasting**.

Bibliography

[1] *About IARPA*. URL: https://www.iarpa.gov/index.php/about-iarpa.

[2] S. D. et al. Anthony. *2018 Corporate Longevity Forecast: Creative Destruction is Accelerating*. URL: https://www.innosight.com/insight/creative-destruction/.

[3] Pavel Atanasov et al. "Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls". In: *Management Science* 63.3 (2017), pp. 691–706. DOI: 10.1287/mnsc.2015.2374.

[4] Glenn W. Brier. "Verification Of Forecasts Expressed In Terms Of Probability". In: *Monthly Weather Review* 78.1 (1950), pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2.

[5] National Research Council. *Intelligence Analysis for Tomorrow: Advances from the Behavioral and Social Sciences*. Washington, DC: The National Academies Press, 2011. ISBN: 978-0-309-16342-2. DOI: 10.17226/13040. URL: https://www.nap.edu/catalog/13040/intelligence-analysis-for-tomorrow-advances-from-the-behavioral-and-social.

[6] Bent Flyvbjerg. "What You Should Know About Megaprojects and Why: An Overview". In: *Project Management Journal* 45 (Feb. 2014). DOI: 10.1002/pmj.21409..

[7] A. Gawande. *The Checklist Manifesto*. Picador, 2011. ISBN: 9780312430009. URL: http://atulgawande.com/book/the-checklist-manifesto/.

[8] Peter Hall. *Great Planning Disasters: With a new introduction*. University of California Press, 1980. ISBN: 9780520046078. URL: http://www.jstor.org/stable/10.1525/j.ctt1ppx64.

[9] D. W. Hubbard. *How To Measure Anything, Third Edition*. Wiley, 2014. ISBN: 9781118539279. URL: https://www.howtomeasureanything.com/.

[10] *Iraq's Continuing Programs for Weapons of Mass Destruction*. 2002. URL: https://fas.org/irp/cia/product/iraq-wmd-nie.pdf.

[11] D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011. ISBN: 978-0374275631. URL: https://us.macmillan.com/books/9781429969352.

[12] J. F. Magee. "Decision Trees for Decision Making". In: (). URL: https://hbr.org/1964/07/decision-trees-for-decision-making.

[13] J. Z. Muller. *The Tyranny of Metrics*. Princeton University Press, 2018. ISBN: 0691174954. URL: https://press.princeton.edu/titles/11218.html.

[14] Z. Nation. *Perceptions of Probability and Numbers*. URL: https://github.com/zonination/perceptions.

[15] P. E. Tetlock and D. Gardner. *Superforecasting*. Broadway Books, 2015. ISBN: 9780804136716. URL: https://wdp.wharton.upenn.edu/book/superforecasting/.