# CRITICAL THINKING IN PSYCHOLOGY

*Second Edition*

EDITED BY

ROBERT J. STERNBERG

*Cornell University*

DIANE F. HALPERN

*Claremont McKenna College*

CAMBRIDGE
UNIVERSITY PRESS

CHAPTER 9

# *When All Is Just a Click Away*
# *Is Critical Thinking Obsolete in the Digital Age?*

*Gerd Gigerenzer*
Max Planck Institute for Human Development, Berlin

## Introduction

Algorithms, so we are told, are more accurate than humans – and cheaper and less biased to boot. Proof of that claim appears to be mounting. Already back in 1997, Deep Blue defeated the world chess champion Garry Kasparov. Today, commercial chess programs play at a similar level. IBM's Watson won on the quiz show Jeopardy against two of its best contestants. Google's Alpha Go beat a world champion at Go. Algorithms are used by the US criminal justice system to predict whether a defendant is likely to reoffend, by the Chinese government to give to every citizen a social credit score that measures their trustworthiness, and by online dating sites to find the ideal romantic match. On the basis of users' search terms, Google Flu Trends predicts the spread of the flu and Microsoft predicts whether a user has pancreatic cancer. Are human judgment, intuition, and expertise, like old typewriters, now redundant, to be replaced by computers? Can humankind eventually cease to think and reflect, and just click and "like"?

I do not think so. In the digital age, we need critical thinking, perhaps more of it than ever before. Yet there is a problem.

Time for critical thinking has become a scarce resource. Sharing, posting, and checking one's phone every couple of minutes eats up time and distracts rather than focuses attention. In fact, digital devices have been programmed to create the need to be constantly connected, intensifying the "fear of missing out" in digital social life. Compared to earlier generations, the age group "iGen" born after 1995 worries more, feels lonelier, is more depressed, and experiences unprecedented levels of social anxiety (Twenge, 2017). "I am constantly worried what others think of my posts and pictures," a 20-year-old English user explained (Royal Society for Public Health, 2017, p. 8). Feeling remote-controlled by others is not the most fertile ground for critical thinking.

## What Is Critical Thinking?

To paraphrase the Enlightenment philosopher Immanuel Kant, critical thinking is the emergence from one's self-imposed nonage. Nonage is the inability to use one's mind without another's guidance. This inability is self-imposed if its cause lies not in the limits of the mind but in the lack of courage to use it. In Kant's (1784) lucid phrase, "Dare to know."

Thus, critical thinking requires knowledge *and* courage. Neither is sufficient in itself. Knowing but not daring is smart but cowardly. Daring without knowing is bold but unwise – it makes one vulnerable to propaganda and fake news. Fake news is not an invention of the digital age; they are old bedfellows. Like a tsunami, they follow great technological breakthroughs that create new communication channels. For instance, when Johannes Gutenberg invented the printing press in 1439, Europe witnessed a wave of fake news on printed broadsides and pamphlets.

Nor did digital media fashion the art of fooling others; what it has done is amplified its reach. Have you been contacted by a distraught widow in Nigeria to help transfer her multimillion heritage to your safe bank account, asking you to send money to cover customs and taxes in return of a promised $1 million? Or have you received an email that you are the lucky winner of $100,000 in an overseas lottery? How could anyone fall for that? Yet people are taken in by hope and greed. As one British victim explained: "That amount of money gives you dreams, and you don't want them taken from you" (Lea, Fischer, & Evans, 2009, p. 42). In the UK alone, every year some three million adults fall victim to mass marketed scams and together lose about £3.5 billion (Office of Fair Trading, 2007). It is not that these victims never noticed anything fishy about the offer. But most acted against their gut feeling because the size of the possible reward was so large that they wanted to give it a try.

People fail to think critically for two reasons. The first is the one just illustrated – that desire trumps critical thinking (Gigerenzer, 2017a). Texting while driving is a similar instance. Most drivers who text know they are risking the lives of others yet cannot stop. As an 18-year-old from Connecticut explained, "I know I should, but it's not going to happen. If I get a Facebook message or something posted on my wall ... I have to see it. I have to" (Turkle, 2011, p. 171). An estimated ten people are killed every day in the US by distracted drivers who cannot ignore the siren's call of their smartphone. Their lack of digital self-control kills more people in the US than terrorists do.

A second reason is that many have not learned the skills for critical thinking. Skills include knowing what questions to ask, how to evaluate numbers, and how to tell fake news from facts. These cognitive skills

combine to form *digital risk literacy* (Gigerenzer, 2014a). They are part of a more general ability called *risk literacy*, which is necessary for enjoying a modern technological world without being harmed in terms of one's health, wealth, or happiness. Both of these reasons go hand in hand: Desire gets in the way of thinking, and the inability to think critically allows others to take advantage of one's desires.

In this chapter, I will provide tools for digital risk literacy. I begin with a fundamental desire and an algorithm that promises to fulfill the hope.

## Find True Love

A good-looking young woman with long hair smiles from a large poster at the entrance to a supermarket. A second poster shows a handsome young man with an attractive three-day beard, smiling as well. Next to their faces is the name of one of the largest European online dating sites, Parship. Both posters prominently display the same catchphrase:

*Every 11 Minutes, a Single Falls in Love*

Eleven minutes – quite a claim. If true, then happiness is just a click away and within months you might be happily attached to your dream partner. Millions of people, mostly with above average education and income, appear to think so and pay several hundred euros for a six-month premium membership, hoping to be one of those who fall in love every eleven minutes with the help of the algorithm.

But we need to think more carefully. Every eleven minutes, someone falls in love. That would be fantastic news if the site had only a few hundred customers. Yet Parship has millions of members. Let us do simple math. One person who falls in love every eleven minutes means about six per hour, which makes 144 per day – assuming that users are active 24/7 on the website. In a year, that makes 52,560 (144 × 365) users. If Parship has one million premium members, only roughly 5 percent of them fall in love within a year. In other words, you may have to wait (and pay) a long time to fall in love, if you do not cancel your membership beforehand (Bauer, Gigerenzer, & Krämer, 2015). Those are the hard facts behind the persuasive catchphrase.

Consistent with this calculation, in 1,500 evaluations of five online dating sites, including Parship, none received an average rating of good from their customers. Only 5 percent said that their search was successful; the rest had quit or are still looking (DISQ, 2017).

Now we know what "every 11 minutes" means. But what about the second part of the catchphrase: "a single falls in love"? It takes two to fall in love to

make a couple. How was falling in love determined? It turns out that every eleven minutes a premium Parship member quit who no longer wished to pay for the service, and when asked why, clicked the "fell in love" button. Whether true love was in play or just a handy excuse, we do not know.

Dating sites are the digital version of the time-honored profession of matchmakers. For Parship, the love formula was developed by a psychologist from the University of Hamburg, based on a psychoanalytic personality test using Rorschach ink-blots and questions about habits such as "Do you sleep with open or closed windows?" and "Do you like to cook?" These questions are designed to dig into actual behavior and interests, as opposed to online dating sites that feature visual and verbal self-presentation.

Deceptive numbers make hope spring eternal and fuel the growth of the online dating market. Going to a party, meeting colleagues and friends, or attending a dance class might be a faster route to happiness. Here is the first principle of being digitally critical:

> If an advertised fact impresses you, be skeptical. The marketing department may have made a mountain out of a molehill.

Dating has fundamentally changed in the digital age. From the 1970s to the 1990s, about 80 percent to 90 percent of 12th-graders in the US used to go out on dates. By 2010, only about 70 percent did so, and by 2015 a mere 55 percent (Twenge, 2017). At the same time, the teen birth rate went down. Among 18- to 19-year-olds, it dropped from around 80 to 90 per 1,000 between 1980 and 2000 to 40 by 2015. The more time spent on social media, the less time for physical dating and, as a consequence, a substantial reduction in teenage pregnancies and births.

Waiting to fall in love via a dating site for years may be a waste of time and a mixed experience. But there are more dangerous risks to not thinking critically.

## Online HIV Tests

The first of December is World AIDS Day, reminding everyone to "know your status" (UNAIDS, 2018). AIDS (acquired immunodeficiency syndrome) is a progressive failure of the immune system caused by infection with HIV (human immunodeficiency virus).[1] By infecting vital cells in the immune system, HIV allows life-threatening infections and cancers to

[1] "HIV" here refers to HIV-1, which is the main family of HIV and accounts for about 95 percent of infections worldwide. HIV-2 occurs primarily in West Africa.

thrive. An HIV infection can occur through the transfer of bodily fluids such as blood, semen, vaginal fluids, and breast milk; without treatment, the average survival after infection is about ten years. Even today, people infected with HIV are subject to stigma and discrimination, which is one reason why those with risky behavior tend to avoid knowing their status.

Screening for HIV involves testing ordinary people who do not have symptoms and are in no risk group. Blood banks screen potential donors, immigration officers screen immigrants, and armed forces screen recruits and personnel on active duty, all compulsorily. The US Preventive Services Task Force recommends that clinicians screen for HIV infection in all adolescents and adults aged 15 to 65 years, including all pregnant women (US Preventive Services Task Force, 2019b). In the analog past, a client made an appointment with a doctor, but in the digital age one can order a rapid HIV test online. Self-testing may reach people who are uncomfortable asking a doctor, but it requires understanding what a test result, positive or negative, actually means.

Imagine a woman, newly married and pregnant. She has no reason to assume that she is infected with HIV but follows the general recommendation to know her status. After ordering a rapid test online, she reads through the instructions, punctures the tip of her left index finger, and extracts a large drop of blood. Then she releases the drop into a small test bottle, sets a timer, and waits ten minutes for the result. The result is not what she thought. It is positive. She cannot believe her eyes.

What does a positive result mean? The instructions accompanying the rapid test say: "You are likely HIV-positive."[2] But how likely is "likely"? After all, it depends on the accuracy of the test. The instructions provide exactly two numbers:

> Specificity: 99.8%
> Sensitivity: 100%

The *specificity* of a test is the probability that a person tests negative if not infected with HIV. Here, it means that in 99.8 percent of all people without infection, the test result is correct. That in turn means that the result is false positive in only 0.2 percent of these cases. This is called the false positive rate (or, false alarm rate). The *sensitivity* is the probability that a person tests positive if infected. Here, the instructions say that the sensitivity is 100 percent. These impressive numbers suggest that it is practically certain that the woman is infected with HIV. She might now

[2] www.autotest-sante.com

ponder how to tell the news to her husband and family. In similar situations, people have considered or actually committed suicide in order to avoid living through social humiliation and the dreaded physical consequences of the disease (Gigerenzer, 2002).

But think for a moment. What is the probability that a person with a positive screening test result is actually infected with HIV? The answer is neither 99.8 percent nor 100 percent. Nor is it revealed or explained how to find it in the instruction sheet.

Consider women who do not practice risky behavior, for whom the frequency of (undiagnosed) HIV infection is about one in every 10,000, which is the case for female US blood donors (Gigerenzer, 2013). Now think of 10,000 of these women who take the test. We expect that one is infected, and that she will test positive (100 percent sensitivity). Of the 9,999 women who are not infected, however, 20 are also expected to test positive. This follows from the false positive rate of 0.2 percent. Thus, we expect that 21 women will test positive, and that only one of them is actually infected. In plain words, it is more likely that the woman is not infected after testing positive for HIV with a rapid test. Thus, the phrase in the instruction sheet "you are likely infected" actually means only one out of 21, or about 5 percent.

A general principle of critical thinking underlies this example. There are two ways to think about all test results. One is confusing for most people, but it is the preferred one in medical education and also in instructions to rapid HIV tests. Here, the information is communicated in *conditional probabilities*, such as sensitivity and specificity. A conditional probability is a probability of $x$ given $y$, such as of a positive test given an infection. The tree on the left in Figure 9.1 shows four conditional probabilities in its bottom branches.

Why do conditional probabilities make it difficult to understand test results? This can be seen from Equation 1. The probability $p$ of a hypothesis $H$ (such as HIV infection) given data $D$ (such as a positive test result) can be calculated by using *Bayes' rule*:

$$p(H|D) = p(H)p(D|H)/[p(H)p(D|H) + p(-H)p(D|-H)] \quad (1)$$

where $p(H|D)$ is the posterior probability, $p(H)$ is the prior probability, $p(D|H)$ is the conditional probability of $D$ given $H$, and $p(D| - H)$ is the conditional probability of $D$ given that the hypothesis is not true $(-H)$. Using the values in Figure 9.1 (left tree), one gets

$$p(HIV|test\ positive) = 0.0001 \times 1.0/(0.0001 \times 1.0 + 0.9999 \times 0.002) \approx 1/21$$

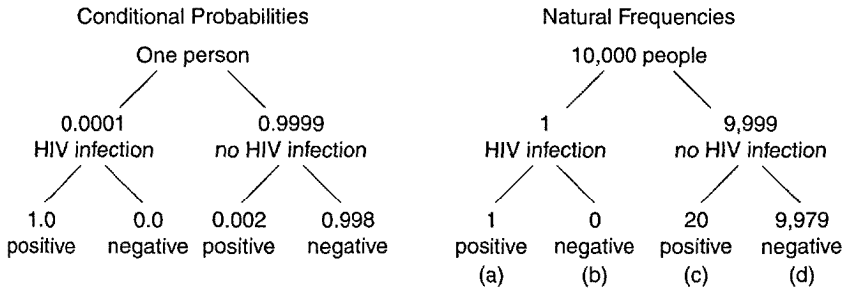


Figure 9.1 Probability of HIV infection when test is positive
What is the probability of being infected by HIV if the rapid test is positive? Conditional probabilities (left tree) tend to confuse people, while natural frequencies (right tree) aid comprehension.

This is how Bayes' rule is taught in textbooks in psychology and medicine. If your mind fogs over after reading this, you are not alone. But there is another option.

This second method is to use natural frequencies (Gigerenzer & Hoffrage, 1995, 1999). The tree on the right-hand side of Figure 9.1 does that. The four frequencies at the bottom of the tree are called *natural frequencies* because – unlike conditional probabilities or relative frequencies – they are not conditionalized (that is, they add up to the total sample size, here 10,000). Natural frequencies simplify understanding test results:

$$p(H|D) = a/(a+c) \qquad (2)$$

Equation 2 is also a version of Bayes' rule, but applies to natural frequencies. All that needs to be done is to take the number of people with positive test *and* infection ($a$), and divide it by the number of people who have a positive test ($a+c$). In the case of the HIV rapid tests, the result is one divided by 21, as we have seen. Natural frequencies facilitate critical thinking. One simply begins with a sample of people, such as 10,000, and translates the probabilities into natural frequencies, as explained above. It is a general technique that can be learned in a few hours.

## Do AIDS Counselors Understand HIV Test Results?

Although many laypeople may not be able to understand the results of medical tests, one might expect that professional AIDS counselors are trained to understand the results of the HIV tests they order. Is that so? To find out, one of my male students, heterosexual and age 25, bravely went

undercover to 20 public health centers to take an HIV test at each and ask the counselors what the test results mean (Gigerenzer, Hoffrage, & Ebert, 1998). It was not an easy study; he could visit only two centers in sequence and then had to wait until the needle traces on both arms healed so that he could continue without being suspected of being a drug addict and thus as belonging to a high-risk group. The centers used the full sequence of ELISA and Western blot tests, which at the time had a probability of about 50 percent that a person with a positive screening test is actually infected (as opposed to only about 5 percent with a rapid test). In the mandatory pre-test counseling, the student asked the questions that everyone should ask, such as "Could I test positive even if I am not infected with HIV? And if so, how often does this happen?" And "If I test positive, would that mean that I am infected for sure? If not, how likely is the infection?"

To our surprise, most professional counselors had no idea. Thirteen counselors wrongly asserted that false positives would never occur, and ten (out of eighteen; two refused to answer the client's question) asserted that a positive result would mean that the client is infected with absolute certainty. Moreover, among the minority of counselors who understood that false positives occur, most were confused about conditional probabilities. For instance, one counselor explained that the test has a sensitivity and specificity of 99.8 percent, and then wrongly asserted that the probability that the client is infected after testing positive is also 99.8 percent (Gigerenzer, 2002).

To make the public health centers aware of this problem, we provided feedback to all counseling centers in Germany after the study. Seventeen years later, we checked whether counseling had improved, using a representative sample of 32 public health centers (two in each of the sixteen German federal states). Another one of my male students went undercover to these centers (Prinz et al., 2015). Although the HIV tests had improved over the years (particularly the false positive rate), the counselors' risk literacy had not. Most still did not understand what a positive test result meant. As in the first study, about half wrongly asserted that false positives never occur and that a positive test means that the client is infected with certainty. Those who did not share this illusion of certainty were confused about conditional probabilities and had not learned to use natural frequencies. Only one counselor could correctly explain the chances.

The instruction that comes with all rapid tests for HIV recommends contacting a physician in the event of testing positive. This is a very good recommendation, but as in the case of the HIV counselors, there is also a high chance that the physician does not understand what a positive test means. Studies have shown that the majority of doctors in the US,

Germany, and elsewhere do not understand health statistics (Gigerenzer et al., 2007). Thus, there is another general lesson in critical thinking:

> Do not assume that health professionals understand health statistics. Be skeptical and have the courage to think about the numbers yourself.

The problem is not that human brains have engrained biases that make it difficult to think statistically, as it is sometimes suggested in the psychological literature. The problem is that medical schools fail to teach tools to understand health statistics, such as using natural frequencies instead of conditional probabilities. In continuing education workshops on statistical thinking, physicians and medical students have been shown to learn quickly when provided the proper tools (e.g. Jenny, Keller, & Gigerenzer, 2018).

## From Query to Cancer

Imagine you are surfing on your search engine when suddenly a warning pops up: "Attention! There are signs that you might have pancreatic cancer. Please visit your doctor immediately." You might be understandably in a state of shock and panic – pancreatic cancer is one of the deadliest cancers, a fast killer with no known cure.

The attempt to diagnose cancer through search queries is not science fiction. Web search queries could predict pancreatic adenocarcinoma, Microsoft researchers have argued (Paparrizos, White, & Horvitz, 2016). They analyzed queries by 6.4 million users of Microsoft's search engine Bing and identified those suggestive of a recent diagnosis, such as "I was told I have pancreatic cancer, what to expect?" Then the researchers looked for queries these users had entered months before, indicating symptoms or risk factors, such as blood clot and alcoholism. They concluded that their algorithm "can identify 5% to 15% of cases, while preserving extremely low false-positive rates (0.00001 to 0.0001)," and that "this screening capability could increase 5-year survival" (Paparrizos, White, & Horvitz, 2016, pp. 1, 7). The *New York Times* (Markoff, 2016) echoed this potential breakthrough: "The study suggests that early screening can increase the five-year survival rate of pancreatic patients to 5 to 7 percent, from just 3 percent."

It seems that Microsoft researchers have found an effective early diagnosis system that produces almost no false alarms and saves lives to boot, which would be huge progress over previous attempts with biomarkers and imaging. The incredibly low false-alarm rate means that if the warning pops up, it is practically certain that the user actually has pancreatic cancer.

Or does it? Think for a moment. A false alarm rate of 0.0001 means that out of every 10,000 users without pancreatic cancer, one is expected to get a false alarm. But that is not the relevant probability that a person has pancreatic cancer if the window popped up. This probability was also missing in the original article, similar to the missing probability in the instructions to the rapid HIV tests. Let us make a simple back-of-the-envelope calculation. Assume 100,000 users, of which ten have undetected pancreatic cancer.[3] With a sensitivity of 10 percent (the average of 5 percent and 15 percent), one expects that one user with pancreatic cancer correctly tests positive while the other nine are missed. Given a false-alarm rate of one in 10,000, we expect ten users to test positive although they do not have cancer. Thus, we expect a total of eleven users to test positive, of which ten do not have pancreatic cancer (Gigerenzer, 2017b). This example illustrates a general point:

> Even with a small false-positive rate, the proportion of false alarms among people who test positive can be quite high if the disease is rare.

## Don't Be Fooled by Survival Rates

Although neither the original study nor the *New York Times* estimated the posterior probability, both featured a different statistic: an increase in the five-year survival rate. If early detection increases the survival rate, it surely then saves lives.

If you agree with that assumption, you have been taken in. Survival rates are a popular tool used to mislead people about the benefit of screening. In spite of the term *survival*, these rates say absolutely nothing about whether screening reduces mortality and prolongs lives. In fact, the correlation between increases in survival rates and decreases in mortality rates is zero ($r = 0.0$) for the 20 most common solid tumors over the past 50 years (Welch, Schwartz, & Woloshin, 2000).

Why do survival rates tell us nothing about reduction in mortality through screening? One reason for this is the *lead-time bias*. Figure 9.2 shows two groups of people, all with invasive cancer, such as prostate cancer, and all of whom die at age 70. The group at the top does not participate in screening, and their cancer is detected at age 67. For them, the five-year survival rate is zero. The bottom group participates in screening and their

[3] This is the estimate given by lead author Paparrizos; see https://www.cs.columbia.edu/2016/web-searches-as-an-early-warning-system-for-pancreatic-cancer

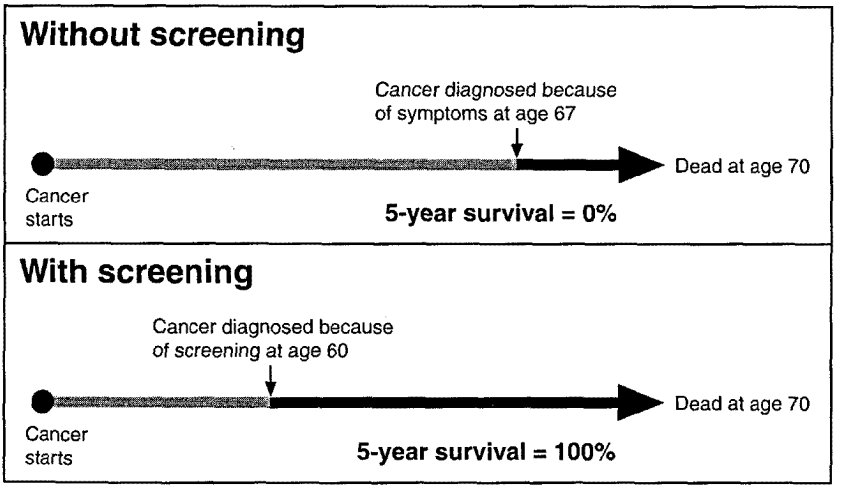


*Figure 9.2* Lead time bias
In screening, an increase in survival rates does not mean that lives are saved or that lives are prolonged. Source: Gigerenzer, 2014a, p. 189

cancer is detected earlier, at age 60. In that case, the five-year survival rate is 100 percent. Despite the 100 percent increase in survival rates due to screening, not a single life was saved or prolonged.

Early detection implies that the time of diagnosis is earlier. That in itself leads to higher five-year survival rates, even if there is no known cure and patients do not live any longer (Gigerenzer, 2014b). Although this fact has been known for decades, reporting survival instead of mortality is still a widely used tool that misleads the public about the benefits of cancer screening. The multi-billion business of cancer screening feeds into the resulting unwarranted hopes. The US Preventive Services Task Force (2013, 2019a) recommends against routine screening for pancreatic cancer because there is no evidence that it reduces mortality. But screening has the potential for significant harm due to surgery or chemotherapy.

Do doctors understand the difference between survival rates and mortality rates? A representative study of more than 400 primary-care physicians in the US showed that the majority – three quarters – did not (Wegwarth et al., 2012). They were mistakenly impressed by improvements in five-year survival due to cancer screening. Politicians have been equally unaware of the difference between survival and mortality rates. After being diagnosed with prostate cancer, former mayor of New York City Rudi

Giuliani praised the US medical system because national survival rates from prostate cancer are 82 percent, compared with only 44 percent under socialized medicine in England. However, the *mortality* rates from prostate cancer were about the same in both countries (Gigerenzer et al., 2007). Similarly, Prime Minister Tony Blair complained about the lower survival rates for colon cancer in the UK than in the US and set a target of increasing them by 20 percent over the following ten years. In 2018, Prime Minister Theresa May repeated the same argument, apparently sharing Giuliani's and Blair's misunderstanding that the higher five-year survival rates in the US mean that people there live longer. The sad truth is that to date, cancer screening – be it for pancreatic, breast, or prostate screening – has never been shown to save lives, that is, to reduce *total* mortality. Instead, it has been shown to lead to harms by unnecessary biopsies, surgery, or radiation (Prasad, Lenzer, & Newman, 2016). In contrast, behavior change such as not smoking, drinking less alcohol, walking, and sports can be effective preventive measures.

In the *New York Times*, one of the authors of the Microsoft study, Dr. Horvitz, stated that he hopes to stimulate quite a bit of interesting conversation about big data and health care. In my view, the first step toward demonstrating the clinical usefulness of big data analytics would be to communicate fewer misleading statistics and more transparency.

## Complex Algorithms or Simple Heuristics?

The advent of big data analytics suggests that more is always better. If a phenomenon is fairly stable so that fine-tuning on data pays, that is likely true. Accordingly, the successes of artificial intelligence that incorporate big data are predominantly in stable situations, such as in the games of chess and Go, where the same rules hold today and tomorrow, and in face and speech recognition. Yet more is not always better in situations of uncertainty, which are unstable and uncertain, such as in predicting the future. Here, fine-tuning on past data can increase prediction error. To counter prediction error, "less can be more." Under uncertainty, simple rules, also known as *heuristics*, can often predict better than complex algorithms do (Gigerenzer & Brighton, 2009). In addition, simplicity also allows for transparency, which is a value per se in an age where nontransparent algorithms influence sensitive decisions, such as about one's creditworthiness, where those who are denied a loan are unable to find out why they were rated negatively. Nevertheless, the belief that

complex algorithms are more accurate and less prejudiced than humans has also made its way into the criminal system.

### *Predictive Policing*

In some countries, including the US, algorithms are used by the criminal system to predict where the next crimes are most likely to occur, and whether a defendant is likely to fail to appear at a court hearing, to reoffend at some point in the future, or to commit a violent crime. For instance, the COMPAS (Correctional Offender Managing Profile for Alternative Sanctions) algorithm has been used in US courts to predict the probability of recidivism for more than one million defendants, influencing pretrial, parole, and sentencing decisions. The commercial algorithm uses 137 features about a defendant and the past criminal record to predict whether the defendant will commit a misdemeanor or felony within the next two years.

In 2013, Eric Loomis drove a car that had been used in a shooting, was arrested, and pleaded guilty to eluding an officer. In order to determine his sentence, the judge looked not only at his criminal record but also at his COMPAS score, which classified him as at high risk of reoffending. The judge sentenced Loomis to six years. Neither the defendant nor the judge could understand how this risk was determined – the algorithm is a business secret. Loomis appealed on the grounds that the judge violated due process by relying on a secretive algorithm. However, the Wisconsin Supreme Court ruled against him, albeit with a recommendation of caution and skepticism in the use of such algorithms (Yong, 2018).

But what are the grounds for skepticism? Are not algorithms the most impartial judges, less biased and more accurate than humans? This seems to have been an implicit assumption of the people working in the criminal justice system. Certainly, few appear to have considered the question until a ProPublica study showed that the COMPAS assessment had a racial bias. And with respect to accuracy, a subsequent study showed that in predicting recidivism, COMPAS fared no better than Amazon Turk workers without any previous experience with recidivism who were paid $1 for quickly assessing the risk of 50 defendants (Dressel & Farid, 2018).

Assessing recidivism is fraught with uncertainty. As mentioned above, in uncertain situations, simple rules can perform better than complex algorithms. Moreover, simple risk assessment tools are transparent, which allows judges and defendants to understand how they work. Simple rules rely on just a few variables that are known to be linked to what one wants to know, here, recidivism. Dressel and Farid (2018) reported that a simple heuristic that relies

on only two variables can match the performance of the COMPAS algorithm and is transparent to boot. The two variables in the simple rule are:

1. age of the defendant and
2. number of previous convictions.

None of the other more than one hundred variables that COMPAS measures appears to add to the quality of the prediction. The general point is that under high uncertainty, fine-tuning with many variables is not the best approach; identifying the few most powerful variables is more promising (Gigerenzer & Gaissmaier, 2011).

The COMPAS algorithm is based on some one hundred data points. But what about big data with millions of data points? Here, more data should be always better, so one might think. Is that true? Consider the celebrated showcase of big data analytics.

### *Google Flu Trends*

In 2008, Google launched Google Flu Trends to predict the spread of flu and influenza-related diseases. The idea behind the algorithm is that users infected by the flu will enter queries into Google's search engine in order to diagnose their symptoms and look for remedies. To develop the algorithm, about 50 million search terms and their correlations with flu-related physician visits were analyzed, and 45 of these terms were chosen. In Google's initial study (Ginsberg et al., 2009), the algorithm was trained on data from the years 2003–2007 and tested on data from 2007–2008. It performed well and was hailed a huge success of big data predictive analytics in bestsellers such as *Big Data* (Meyer-Schoenberger & Cukier, 2013).

In 2009, however, the swine flu broke out at an untypical time of the year: The first cases were reported in March and the outbreak peaked in October. Google Flu Trends failed to predict this outbreak (Olson et al., 2013). Even after the Google engineers revised the algorithm, it continued to be inaccurate. For instance, from August 2011 to September 2013, the algorithm overestimated the proportion of flu-related physician visits in 100 out of 108 weeks (Lazer, 2014). In 2015, Google Flu Trends was shut down, without any fanfare and attention from the media that had hailed it only few years before.

Why did Google Flu Trends fail? First, the algorithm had learned that flu levels are typically high in the winter and low in the summer – but the swine flu appeared out of season. The future was not like the past. Second,

big data algorithms "think" in terms of correlations, not causes. However, the wide media coverage of the swine flu caused flu-related searches by people who had no symptoms but were simply curious. The algorithm cannot distinguish between these two causes for searches – being sick or curious – and overestimated flu-related doctor visits. When the algorithm was built, the engineers had to eliminate by hand search terms that had high correlations but were obviously not causally related to the flu. One example is *high school basketball*, whose season coincides with the flu season (Ginsberg et al., 2009).

Finally, the Google Flu Trends algorithm might have been unnecessarily complex. In an unstable world, less is often more (Gigerenzer & Gaissmaier, 2011). The original study did not test simple heuristics as candidates. Later it was shown, however, that regressions with only a few variables were more accurate at predicting flu than Google Flu Trends (Lazer et al., 2014). Finally, Katsikopolous, Simsek, Buckman, and Gigerenzer (in press) reanalyzed Google's original data and showed that a simple heuristic that relies on a single, easy-to-access cue performed better than an analysis of 50 million search terms at predicting flu-related doctor visits:

> *Recency heuristic:* Predict that the number of flu-related doctor visits is equal to the number of visits two weeks ago.

The number of flu-related doctor visits one week ago was not available; therefore, the value of two weeks ago was used. This heuristic predicts better than Google Flu Trends and has an error of only .33, compared with .49 for the Google algorithm. No big data analytics are required. Less information can be more effective.

In health care, big data analytics are known for hyperbolic claims. After IBM's computer program Watson won on Jeopardy, it was marketed as the future of health and finance. You may have seen IBM Watson commercials where what looks like a sentient box interacts with Bob Dylan, Serena Williams, and other celebrities. IBM's boldest promises have been about the ways it will have an impact on health care. Hospitals pay a per-patient fee for Watson's services, ranging between $200 and $1,000 per patient. M. D. Anderson, one of the most respected cancer centers in the US, announced collaboration with IBM in 2014 to develop personalized cancer diagnostics and treatment. After spending $62 million, M. D. Anderson found out that the software cannot do what the aggressive marketing department of IBM claimed it could do, and dismissed "Dr. Watson." IBM has not published any scientific papers demonstrating the validity of the technology for physicians and patients. Moreover, if Watson is able to

make better financial investments, IBM should not be in the financial troubles it has faced over the past years. "IBM Watson is the Donald Trump of the AI industry – outlandish claims that aren't backed by credible data," said Oren Etzioni, CEO of the Allen Institute for AI and former computer science professor (Brown, 2017).

## Critical Thinking in the Digital Age

These case studies – finding true love, understanding HIV test results, and predicting cancer, recidivism, and flu – illustrate the general need for critical thinking. At issue is the courage to use one's own mind without others' guidance, and to think through the numbers and claims. Yet critical thinking demands not only courage, but also routine skills. Let me review some the general tools of critical thinking outlined in this chapter, together with new problems for practice.

1. *If an advertised fact impresses you, be skeptical. The marketing department may have made a mountain out of a molehill.* The eleven minutes for falling in love is not the only message that easily misleads us because we like to hear it. The same skepticism is warranted about messages from the press. If you enjoy jogging, you might be receptive to this media headline about a study on longevity (Lee et al., 2017): "For every hour of daily jogging, you live 7 hours longer" (Bauer, Krämer, & Gigerenzer, 2017).

To invest one hour and gain seven – that is a sound investment. But pause to think for a moment. If that claim were true, we could literally run ourselves into immortality. Jogging, say, four hours a day would mean extending one's life by 28 hours. That is more than a 24-hour day, and thus our life expectancy would increase each day. Clearly something is wrong with the media report. Indeed, the original study did not make any such claim. What it did claim was that this effect holds for two hours of jogging per week, but not for additional hours (Lee et al., 2017). More precisely, the seven-hour figure was estimated this way: A group of joggers age 44, who run for two hours per week, spend a total of 0.43 years running by the age of 80 and win 2.8 years of increased life expectancy, which is equivalent to about one hour of running for seven hours of living longer. More running is not necessarily better. On the contrary, excessive running can increase the risk of heart disease and shorten one's life.

2. *False alarms occur even with the best tests.* A false alarm occurs when a person tests positive but does not have the disease. Consider this *Nature* report:

> A British man has hit headlines this week with reports that he has recovered from an HIV infection, having tested positive in August 2002 and negative in the ensuing years. If this proves to be true on further testing, 25-year-old Andrew Stimpson, living in London, would be the first person confirmed to have eliminated HIV from his body. (Hopkin, 2005)

Is that case proof of a miraculous recovery? It is more likely proof of a false positive. As we have seen earlier and as correctly noted by the author of the *Nature* report, false positives occur in HIV tests. When the person is tested again, the result is then likely negative. There is no need to assume that a miracle cure has happened.

3. *The probability of having a disease given a positive test result (the positive predictive value) can be most intuitively determined with natural frequencies.* Here is an exercise. Try to solve the question by translating the probabilities into natural frequencies:

> The best HIV tests (neither rapid nor online) are said to have a false-alarm rate of only one in 250,000 people. Assume a sensitivity of 99.8 percent and a rate of undetected HIV infection of one in every 10,000. What is the chance that a random person who tests positive in HIV screening is actually infected?

First off, the answer is not one in 250,000. In order to deal with such a small false-alarm rate, one can build a natural frequency tree similar to that in Figure 9.1. In this case, consider 250,000 low-risk people who are screened for HIV. We expect 25 to be infected and all of them to test positive. Among those without HIV infection, we expect one to test positive. Thus, one out of every 26 who test positive is not HIV-infected (Gigerenzer, 2013). Even with an extremely low false-positive rate, the posterior probability of being HIV-infected given a positive test is not negligible.

4. *Five-year survival rates are irrelevant for measuring the benefit of screening.* Changes in survival rate have no correlation with changes in mortality rates, which is the relevant figure. Nevertheless, several organizations communicate survival rates (but not mortality rates) to impress the public about the benefit of screening. Consider Susan G. Komen, one of the largest, best funded, and most trusted breast cancer organizations in the US. In a promotion of mammography screening, Komen tells women what they should do: "GET SCREENED NOW":

> "LESS TALK. MORE ACTION. Early detection saves lives. The five-year survival rate for breast cancer when caught early is 98 percent. When it's not? 23 percent."

The promotion provides no other information about benefits, and none about harms (Gigerenzer, 2014a). What follows from the reported increase in survival rates for reducing mortality, that is, saving lives?

The correct answer is, as you may recall: Nothing. Although the differences in five-year survival rates look impressive, they do not carry any information about mortality rates. If the aim of Komen is to inform rather than nudge women, it needs to mention the mortality rates. These can be easily explained. About 500,000 women aged 50–70 were studied in randomized trials, half randomly assigned to the screening group and the others to the nonscreening group. About ten years later, among every 1,000 women who did not participate in screening, about five died from breast cancer. And among every 1,000 women who did participate, about four died from breast cancer. Thus, the breast-cancer mortality reduction from five to four in 1,000 equals one in every 1,000, or 0.1 percent. In the Komen promotion, this is presented as a spectacular increase from 23 percent to 98 percent.

One might argue that even one in 1,000 represents a respectable figure of thousands of lives being saved if one looks at the entire population of a country. But even that is not so. The one in 1,000 figure is for breast-cancer mortality alone. The *total* mortality for all cancers as well as the total mortality remained the same among women who participated in screening and those who did not. In other words, in the screening group, one less woman died from breast cancer but this was cancelled out by one additional woman dying from another cancer, possibly due to radiation in screening or therapy or other consequences of treatment. This fact is almost never passed on to women. In plain words, we have no evidence that mammography screening saves lives but evidence that it hurts many women (Gigerenzer, 2014b; Gøtzsche & Jørgensen, 2013). In a world of health care that has sadly become commercialized, where misleading statistics are the rule rather than the exception, it is important to navigate number-based claims with caution.

## The Big Picture

In the British TV series *Black Mirror* (named after the surface of a smart phone or tablet), there is an episode called "Nosedive." It shows a possible future of humanity in which every person has a social score from 1 to 5 stars, similar to the Amazon ratings. With the help of smart contact lenses and face recognition software, everyone can see everyone else's score. Lacie, a young woman with a score of 4.2, is eager to move into a better part of the city. But to get an affordable apartment, she needs a 4.5. Her life centers on

one and only one goal: to improve her score. She tries hard but eventually fails and ends up in jail.

Nosedive is science fiction. But the scenario is becoming reality in China, whose government announced that by 2020, every citizen will receive a Social Credit Score. This score not only measures financial creditworthiness, as the FICO credit risk score in the US does. It also measures a person's social trustworthiness, political compliance, and interpersonal relationships, using hundreds of millions of surveillance cameras, digital footprints, and every possible data source. Crossing the street at a red light means losing points; visiting one's elderly parents increases them. Those with high scores are rewarded perks, as in a frequent-flyer program, while those with low scores are punished. With a low social credit score, one may not be permitted to fly or use bullet trains, or one's children may not be able to visit the best schools. Even one's social circles have an impact: friends with low scores can pull down one's own score.

The social credit system is made possible by digital surveillance systems. Social surveillance is nothing new, but this technology can scale it up to levels never seen before. Its goal is to improve people's moral behavior, eliminate corruption, and create a culture of "sincerity" and "harmony." And the system already appears to bear fruit. Car drivers have become kind to pedestrians, social media users have "unfriended" those with low scores, and people list their score in personal ads to attract better romantic partners. Surveys indicate that the far majority of Chinese citizens are in favor of this system, particularly those with higher education. Many believe that it provides a true alternative to democracy, fostering morality, harmony, and economic growth.

Here is a possible future. First, the Chinese government will have solved the technological problems of collecting, identifying, and integrating all data into one score at some point in the 2020s. In a second phase, the software and surveillance technology will be sold to other countries with similar autocratic outlooks. That may boost moral behavior and economic growth in these countries as well. Third, democratic governments, notoriously slow in decision making, will face the rise of digital one-party systems that make and implement decisions much more quickly and whose citizens trust the government more than is the case in Western democracies. While democratically elected political parties waste time fighting other parties and dealing with secession or belligerent leaders, and voters are manipulated by Facebook ads, trolls, and other means, the new digital single-party system largely eliminates these problems. Whether present democracies can survive this powerful alternative remains to be seen.

When Westerners learn about the Social Credit System, their reactions are usually of distanced repugnance. Absolute surveillance, where else but in China! A democracy would never monitor the daily activities of its citizens – what they buy, where they are, with whom they are, what websites they visit, whether they pay their bills. In the West, however, this is already happening. We rate restaurants, posts, movies, and even doctors on websites open to the public eye. Many happily give away their personal and biometric data for the benefit of small conveniences. "I have nothing to hide" is the usual defense. Like the Chinese government, Google's Eric Schmidt reminds us: "If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place" (Bartiromo, 2009).

Now imagine a commercial system that integrates all these data, evaluates them as positive or negative, and creates a super score. Companies such as Acxiom claim that they have collected up to 3,000 data points for over 700 million people worldwide, including item-level purchase data, health data, criminal records, credit card activity, voter records, and location data. In the West, chances are that commercial companies will also develop a super score to measure citizens' social trustworthiness.

We need to think about where we want to be in twenty years. Should it be a system where people trust in an authority that distributes goodies to those who collect points and punishes those who do not conform? Or should we update our democratic systems from one where the political parties have lowest trust ratings to a more effective and competitive form? If we do not reflect on these issues and instead let technology lead us aimlessly, we might well end up in a commercially driven social credit system. To what greater extent will privacy be eroded? Is democracy becoming a thing of the past? It is time for us to think critically about the future rather than let it be decided by commercial algorithms or autocratic interests.

### Gerd Gigerenzer: How Critical Thinking Has Played a Role in My Own Career

A most useful discovery in my career was the concept of *natural frequencies*, one of the tools for critical thinking described in this article.

From the 1970s to the 90s, most psychological research took it for granted that people cannot reason according to Bayes' rule (see Equation 1) and, more generally, have a hard time thinking in probabilities. David Eddy (1982), for instance, asked physicians about the chance that a woman with a positive screening mammogram actually has breast cancer. Ninety-five of

100 physicians believed the probability is around 75 percent, when it was only 8 percent. One can imagine what unnecessary anxiety and panic these innumerate physicians unintentionally caused. In a pointed statement of two famous psychologists, the conclusion was that the mind "is not a Bayesian at all" (Kahneman & Tversky, 1972).

Because I enjoy reading interdisciplinarily, I came across studies on animal behavior that concluded that bees, bumblebees, birds and bats are good Bayesians. Something must be wrong, I thought. How can animals excel in reasoning and humans not? I then took a closer look at the studies and found that the difference was not in the species, but in the representation of the statistical information. Humans were tested on stated conditional probabilities, whereas animals encountered frequencies from experience. This learning from experience is called natural sampling, and results in natural frequencies (see Figure 9.1). The surprising theoretical result was that natural frequencies simplify Bayes' rule (Equation 2). Ulrich Hoffrage and I showed for the first time that with natural frequencies, people can solve Bayesian problems better than with probabilities (Hoffrage & Gigerenzer, 1995, 1999), while Peter Sedlmeier and I found that a two-hour training in natural frequencies enabled people to solve about 90 percent of problems, a level maintained when tested three months after training (Sedlmeier & Gigerenzer, 2001). That settled the issue: People think the Bayesian way when the information is in natural frequencies – a format they naturally encounter from direct experience – but appear to be inept when confronted with conditional probabilities, which most have never learned at school.

A good theory has fruitful applications, Kurt Lewin once said. To follow his motto, we left the laboratory to test physicians and found that Eddy's results replicate with probabilities only. Natural frequencies, in contrast, facilitate physicians' understanding of what a test result means. We found the same when testing law students, professors of law, and judges, who could not reason with DNA probabilities but succeeded in understanding the numbers when given natural frequencies (Lindsey, Hertwig, & Gigerenzer, 2003). In a next step, I trained some 1,000 physicians in continuing medical education to translate probabilities into natural frequencies, which helped them to get around 80 percent to 90 percent correct answers in subsequent tests (Gigerenzer, 2014a).

Since then, the use of natural frequencies has been recommended by major medical organizations as a tool for risk communication, and they are now a standard concept in evidence-based medicine. Recently, several ministries of education in German states have replaced learning conditional probabilities with natural frequencies in their school curricula. This tool for critical thinking will hopefully lead to a less "mathematically traumatized" generation of students who understand Bayes' rule and are not left mystified.

## Critical Thinking Questions:

1. James checks the weather report on his smartphone. It says "a 30 percent chance of rain tomorrow." But 30 percent of what? He assumes it means that it will rain tomorrow 30 percent of the *time*. But he's not really sure, so asks his friends. Emma believes it will rain tomorrow in 30 percent of the *area*. Luna thinks it will rain on 30 percent of the days for which this prediction has been made. And Leo says it means that three meteorologists believe it will rain, and seven do not. Who is right?
2. You play a lottery where only one out of every 1,000 tickets wins. You buy one ticket. The probability that you will win is _____ percent.
3. A 50-year old woman without symptoms participates in mammography screening and tests positive. She is frightened by the prospect of having breast cancer. The sensitivity of mammography is 90 percent, the false-alarm rate is 9 percent, and the prevalence of undetected breast cancer in this population is 1 percent. What is the chance that the woman actually has breast cancer? (Hint: translate the percentages into natural frequencies)
4. Studies indicate that increasing smartphone use by parents and children is eradicating face-to-face conversation within families. What would be a small set of easy-to-memorize behavioral rules that a family could implement to bring conversation back into their homes?
5. What can you do to become more independent in your thinking from your peers' opinions, experience less fear of missing out, and take the "remote control" of your emotional life through social media back into your own hands?

## Key Terms

**Bayes' rule** A rule for updating the probability of hypotheses in the light of new evidence. For the simple case of a binary hypothesis ($H$ and $not-H$, such as cancer and not cancer) and data $D$ (such as a positive test), the rule is:

$$p(H|D) = p(H)p(D|H)/[p(H)p(D|H) + p(-H)p(D|-H)],$$

where $p(D|H)$ is the posterior probability, $p(H)$ is the prior probability, $p(D|H)$ is the probability of $D$ given $H$, and $p(D|not-H)$ is the probability of $D$ given $not-H$. Many have problems understanding this rule. But there is help. The interesting point is that the calculation of $p(H|D)$ becomes more intuitive when the input is stated in natural frequencies rather than in probabilities. For natural frequencies, the rule is: $p(H|D) = a/(a+b)$ where $a$

is the number of $D$ and $H$ cases, and $b$ the number of $D$ and $-H$ cases. See *Natural frequencies.*

**Conditional probability** The probability that an event $A$ occurs given event $B$, usually written as $p(A|B)$. Conditional probabilities are notoriously misunderstood, and in two different ways. One is to confuse the probability of $A$ *given* $B$ with the probability of $A$ *and* $B$; the other is to confuse the probability of $A$ given $B$ with the probability of $B$ given $A$. One can reduce this confusion by replacing conditional probabilities with natural frequencies. See *Natural frequencies.*

**False-alarm rate (false-positive rate)** The proportion of positive tests among people without the disease is called the false-alarm rate. It is typically expressed as a conditional probability or a percentage. For instance, mammography screening has a false positive rate of 5 to 10 percent depending on age; that is, 5 to 10 percent of women without breast cancer nevertheless receive a positive test result. The false positive rate and the specificity (the probability of a negative result given no disease) of a test add up to 100 percent. The rates of the two errors are dependent: Decreasing the false positive rate of a test increases the false negative rate, and vice versa.

**Heuristic** A rule of thumb, or *heuristic*, is a conscious or unconscious strategy that ignores part of the information to make better judgments. It enables us to make a decision fast, with little search for information but nevertheless with high accuracy. Heuristics are indispensable in a world where not all risks are known ("uncertainty"), while probability theory is sufficient in a world where all risks are known ("risk"). A rational mind needs both sets of tools. The widespread idea that heuristics are always second best and that more information and computation are always better is incorrect.

**Lead time bias** One reason why survival rates are misleading about the benefits of screening. Even if the time of death is not changed by screening – that is, no life is saved or prolonged – early detection advances the time of diagnosis and thus results in increased survival rates.

**Mortality rate** A measure of the benefit of a treatment in terms of lives saved. For instance, to evaluate the benefit of cancer screening, the mortality rates in the screening group and the control group are compared. The difference is the reduction in mortality. In the context of screening, mortality rates, not survival rates, are the relevant statistics. See *Survival rate.*

**Natural frequencies** Frequencies that correspond to the way humans encountered information before the invention of books and probability theory. Unlike probabilities and relative frequencies, they are "raw" observations that have not been normalized with respect to the base rates of the event in question. For instance, a physician has observed 100 persons, ten of whom show a new disease. Of these ten, eight show a symptom, whereas four of the 90 without disease also show the symptom. Breaking these 100 cases down into four numbers (disease and symptom: 8; disease and no symptom: 2; no disease and symptom: 4; no disease and no symptom: 86) results in four natural frequencies: 8, 2, 4, and 86. Natural frequencies facilitate Bayesian inferences. For instance, a physician who observes a new person with the symptom can easily see that the chance that this patient also has the disease is $8/(8 + 4)$, that is, two thirds. This probability is called the *posterior probability.* Natural frequencies help people to "see" the posterior probabilities, whereas conditional probabilities tend to be confusing. See *Bayes' rule.*

**Posterior probability** The probability of a hypothesis after new evidence, that is, the updated prior probability. It can be calculated from the prior probability using *Bayes' rule* and, more intuitively, using *natural frequencies.*

**Sensitivity** The sensitivity of a test is the percentage of individuals who are correctly classified as having the disease. Formally, the sensitivity is the *conditional probability p(positive| disease)* of a positive test result given the disease. The sensitivity and the false negative rate add up to 100 percent. The sensitivity is also called the hit rate.

**Specificity** The specificity of a test is the percentage of individuals who are correctly classified as not having the disease. Formally, the specificity is the *conditional probability p(negative|no disease)* of a negative test result given no disease. The specificity and the false positive rate add up to 100 percent.

**Survival rate** A measure of the benefit of a treatment: five-year survival rate = number of patients diagnosed with cancer who are still alive five years after diagnosis divided by the number of patients diagnosed with cancer. In the context of screening, changes in survival rates are misleading about the benefit because they do not correspond to changes in mortality rates. One of the reasons is lead time bias. Nevertheless, many institutions promote screening on the basis of survival rates. See *Lead time bias.*

## REFERENCES

Bartiromo, M. (2009, December 3). Google CEO Eric Schmidt on privacy. Interview, CNBC. Online video clip. https://www.youtube.com/watch?v=A6e7wfDHzew

Bauer, T. K., Gigerenzer, G., & Krämer, W. (2015, December 21). Unstatistik des Monats: Liebestrunken – Vermittlungsbörse schieß statisches Eigentor [Bad statistic of the month: Dating site makes a statistical foul]. Online. http://www.rwi-essen.de/unstatistik/50

Bauer, T. K., Krämer, W., & Gigerenzer, G. (2017, April 28). Unstatistik des Monats: Eine Stunde joggen, sieben Stunden länger leben [Bad statistic of the month: One hour of jogging adds seven hours of life]. Online. https://www.mpib-berlin.mpg.de/de/presse/dossiers/unstatistik-des-monats/archiv-zur-unstatistik

Brown, J. (2017, August 10) Why everyone is hating on IBM Watson, including the people who helped make it. Gizmodo. Online. https://gizmodo.com/why-everyone-is-hating-on-watson-including-the-people-w-1797510888

Deutsches Institut für Service-Qualität (DISQ) (2017). Kundenbefragung: Online-Partnerbörsen 2017 [Customer survey: Online dating sites]. Online. https://disq.de/2017/20170426-Online-Partnerboersen.html

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*, eaao5580. DOI:10.1126/sciadv.aao5580

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge: Cambridge University Press.

Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you.* New York: Simon & Schuster.

Gigerenzer, G. (2013). HIV screening: Helping clinicians make sense of test results to patients. *British Medical Journal, 347*, f5151. DOI:10.1136/bmj.f5151

Gigerenzer, G. (2014a). *Risk savvy: How to make good decisions.* New York: Penguin.

Gigerenzer, G. (2014b). Breast cancer screening pamphlets mislead women. *British Medical Journal, 348*, g2636. DOI:10.1136/bmj.g2636

Gigerenzer, G. (2017a). Digital risk literacy: Technology needs users who can control it. *Scientific American, 25*. Online. https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence

Gigerenzer, G. (2017b). Can search engine data predict pancreatic cancer? *British Medical Journal, 358*, j3159. DOI:10.1136/bmj.j3159

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1*, 107–143. DOI:10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451–482. DOI:10.1146/annurev-psych-120709-145346

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. W. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 8*, 53–96. DOI:10.1111/j.1539-6053.2008.00033.x

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684–704. DOI:10.1037/0033-295X.102.4.684

Gigerenzer, G., Hoffrage, U., & Ebert, A. (1998). AIDS counseling for low-risk clients. *AIDS CARE, 10*, 197–211. DOI:10.1080/09540129850124451

Gigerenzer, G., & Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: A reply to Lewis & Keren and Mellers & McGraw. *Psychological Review, 106*, 425–430. DOI:10.1037/0033-295X.106.2.425

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014. DOI:10.1038/nature07634

Gøtzsche, P. S., & Jørgensen, K. J. (2013). Screening for breast cancer with mammography (Review). *Cochrane Database of Systematic Reviews, 6*, CD001877. DOI:10.1002/14651858

Hopkin, M. (2005, November 14). Experts urge caution on HIV 'miracle recovery.' *Nature*. DOI:10.1038/051114-3

Jenny, M. A., Keller, N., & Gigerenzer, G. (2018). Assessing minimal medical statistical literacy using the Quick Risk Test: A prospective observational study in Germany. *BMJ Open, 8*:e020847. DOI:10.1136/bmjopen-2017-020847

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430–454.

Kant, I. (1784). Beantwortung der Frage: Was ist Aufklärung? *Berlinische Wochenschrift*, Dezember Heft 481–494.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). Big data. The parable of Google Flu: Traps in big data analysis. *Science, 343*(6176), 1203–1205. DOI:10.1126/science.1248506

Lea, S., Fischer, P., & Evans, K. (2009). The psychology of scams: Provoking and committing errors of judgement. Office of Fair Trading Report No. 1070. Online. https://tinyurl.com/y452jw47

Lee, D., Brellenthin, A. G., Thompson, P. D., Sui, X., Lee, I., & Lavie, C. J. (2017). Running as a key lifestyle medicine for longevity. *Progress in Cardiovascular Diseases, 60*(1), 45–55. DOI:10.1016/j.pcad.2017.03.005

Lindsey, S., Hertwig, R., & Gigerenzer, G. (2003). Communicating statistical DNA evidence. *Jurimetrics, 43*, 147–163.

Markoff, J. (2016, June 7). Microsoft finds cancer clues in search queries. *New York Times*. https://www.nytimes.com/2016/06/08/technology/online-searches-can-identify-cancer-victims-study-finds.html

Meyer-Schönberger, V., & Cukier, K. (2013). *Big data*. London: John Murray.

Olson, D. R., Konty, K. J., Paladini, M., Viboud, C., & Simonsen, L. (2013). Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. *PLoS Computational Biology, 9*(10), e1003256. DOI:10.1371/journal.pcbi.1003256

Office of Fair Trading (2007). Research on impact of mass marketed scams: A summary of research into the impact of scams on UK consumers. Office of Fair Trading Report No. 883. Online. https://tinyurl.com/y4v2g35c

Paparrizos, J., White, R. W., & Horvitz, E. (2016). Screening for pancreatic adenocarcinoma using signals from web search logs: Feasibility study and results. *Journal of Oncology Practice, 12*(8), 737–744. DOI:10.1200/jop.2015.010504

Prasad, V., Lenzer, J., & Newman, D. H. (2016). Why cancer screening has never been shown to "save lives" – and what we can do about it. *BMJ, 352*: h6080. DOI:10.1136/bmj.h6080

Prinz, R., Feufel, M., Gigerenzer, G., & Wegwarth, O. (2015). What counselors tell low-risk clients about HIV test performance. *Current HIV Research, 13*, 369–380. DOI:10.2174/1570162X13666150511125200

Royal Society for Public Health (2017). #StatusOfMind: Social media and young people's mental health and wellbeing. Online. https://tinyurl.com/y990p90j

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130*, 380–400. DOI:10.1037/0096-3445.130.3.380

Turkle, S. (2011). *Alone together*. New York: Basic Books.

Twenge, J. M. (2017). *iGen*. New York: Atria Books.

UNAIDS (2018, September 17) World AIDS Day 2018 theme encourages everyone to know their HIV status. Online. http://www.unaids.org/en/resources/presscentre/featurestories/2018/september/20180917_WAD_theme

US Preventive Services Task Force (2013). Pancreatic cancer: Screening. Final recommendation statement. Online. https://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/pancreatic-cancer-screening

US Preventive Services Task Force (2019a). Pancreatic cancer: Screening. Draft recommendation statement. Online. https://www.uspreventiveservicestaskforce.org/Page/Document/draft-recommendation-statement/pancreatic-cancer-screening1

US Preventive Services Task Force (2019b). Screening for HIV infection: US Preventive Services Task Force recommendation statement. *JAMA, 321*(23), 2326–2336. DOI:10.1001/jama.2019.6587

Wegwarth, O., Schwartz, L. M., Woloshin, S., Gaissmaier, W., & Gigerenzer. G. (2012). Do physicians understand cancer screening statistics? A national survey of primary care physicians. *Annals of Internal Medicine, 156*, 340–349. DOI:10.7326/0003-4819-156-5-201203060-00005

Welch, H. G., Schwartz, L. M., & Woloshin, S. (2000). Are increasing 5-year survival rates evidence of success against cancer? *JAMA, 283*, 2975–2978. DOI:10.1001/jama.283.22.2975

Yong, E. (2018, January 17). A popular algorithm is not better at predicting crimes than random people. *The Atlantic*. https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/55064`6