Material excerpted from
AIQ: How People and Machines are Smarter Together
By Nick Polson and James Scott
St. Martin's Press (May 15, 2018)
https://www.amazon.com/AIQ-People-Machines-Smarter-Together/dp/1250182158/

Act 2: How Effective Is Your Method?

The Egyptians of the Lower Kingdom used a mixture of honey and sodium carbonate. The Mesopotamians preferred acacia leaves and lint. The ancient Persians used elephant dung and cabbages; the Renaissance Europeans, lily root and silkworm gut.

In modern societies, we have it a bit easier. Most people choose condoms or the pill, or they get voluntarily and painlessly sterilized.

Birth control is at least as old as civilization; the big difference from ancient times is that our methods actually work well. Since the 1960s, when effective contraception became widely available, birth rates across the industrialized world have plummeted. Today, some experience with contraception is nearly universal among sexually active adults in rich countries.[9]

We recognize that, for many people, the choice of when to use contraception, and what method to use, cannot be reduced to a single variable.[10] But one important question for everyone is the chance of getting pregnant if you use a particular method. It was with this question in mind that, in 2014, *The New York Times* published an article entitled "How Likely Is It That Birth Control Could Let You Down?"[11] The authors of the article began from a simple premise: the more times you use any method of birth control, the more opportunities there are for it to fail. To put some numbers behind this idea, the authors of the article looked at published data on the 1-year efficacy of 15 popular contraceptive methods. They used this data—together with their own calculations, which we'll describe below—to create a slick interactive chart that purported to show each method's long-term failure rate, out to 10 years.

We have used the same published data to replicate the *Times*'s calculations,[12] using the same methodology employed by the authors of the article, for a subset of nine of these methods. Our calculations, which you can see in Figure 7.1, agree with theirs. Each panel shows a different contraceptive method. The vertical axis shows the *Times*'s estimate for the probability of getting pregnant at least once if you use that method over the long run.

If the numbers in this figure surprise you, you are not alone: the *Times* article shocked a whole lot of people. For example, the article

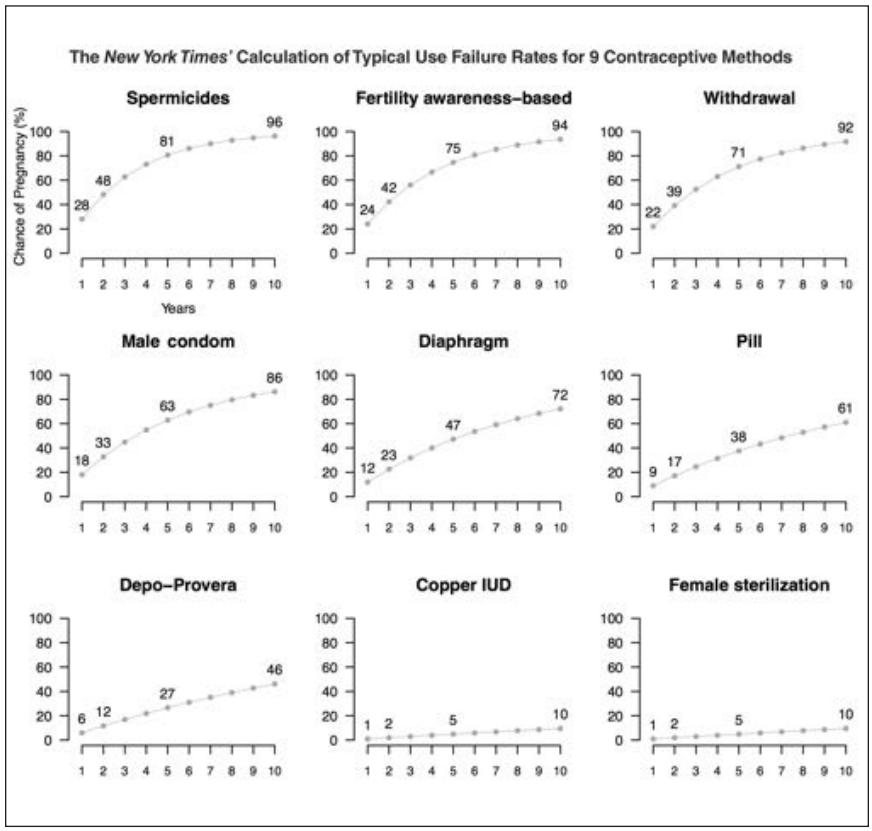

Figure 7.1

stated that the 1-year failure rate for typical pill users was 9%,[§] but that the 10-year failure rate was an alarming 61%. The numbers for the condom were even worse: its 10-year failure rate was shown at 86%. To many people, these numbers seemed astonishingly high, and implied a far greater long-term risk of unplanned pregnancy than they were prepared for. Perhaps as a result, the article quickly went viral on social media—and while it may not have sparked a mass rush to join a convent, it did cause a lot of anxiety among ordinary *Times* readers, many of whom had presumably believed their own contraceptive methods to be more reliable. Even gynecologists, who should know the research

§ "Typical use" does not mean "correct use." If you use the pill exactly as instructed, the failure rate is much lower, at less than 1% per year.

—-1
—0
—+1

just about as well as anyone, went straight online to share the link and to express their alarm.*

A Story Built on Poor Assumptions

But there was a major problem with the *New York Times* article: its putative long-term failure rates have no basis in fact. They're almost surely *way* too high.

It turns out that nobody in the world actually knows the 10-year failure rates for any of these contraceptive methods.[13] For practical reasons, the question just hasn't been studied. Despite this lack of evidence, however, there are strong reasons to believe that, because of poor assumptions, the *Times* article drastically overstated the chance of getting pregnant under long-term use of most contraceptive methods.

Here's how the *Times* calculated each method's purported long-term failure rate. First, they took the one-year "typical-use" failure rate from published research (for example, 9% for the pill). These one-year numbers were originally calculated using data from clinical trials or nationally representative surveys. They were the best estimates available. So far, so good.

Next, the authors used the compounding rule to calculate the probability of a no-pregnancy "winning streak" for several years in a row. Effectively, the *Times* journalists were treating a multiyear stretch of contraceptive use without a pregnancy exactly as if it were a Joe DiMaggio hitting streak, using the same two assumptions we used earlier: independence and constant probability across years.

Let's see an example. Among typical pill users, the probability of successfully avoiding pregnancy in year 1 was 91%. Based on this figure, the *Times* used the compounding rule to calculate the following probabilities:

$$P(\text{no pregnancy through 1 year}) = 0.91$$
$$P(\text{no pregnancy through 2 years}) = (0.91)^2 \approx 0.82$$
$$P(\text{no pregnancy through 3 years}) = (0.91)^3 \approx 0.75.$$

-1—
0—
+1—

* For example, @hricciot: "Shocking—even to a gynecologist like me! #LARCisBest." "LARC" stands for "long-acting reversible contraception"—for example, an IUD.

And so on. By the time you get out to 10 years, the probability of a long "winning streak" starts to look pretty small: about 39%. This implies a 61% chance of at least one pregnancy over a 10-year period of typical pill use.

An Analogy

This calculation, however, has an enormous flaw. To see what it is, let's reason by analogy. Suppose that we conduct a study by recruiting 100 people and giving each of them a coin. These coins have been altered so that 90 of them have heads on both sides, and 10 of them have tails on both sides. Now we have our study participants start flipping their coins. We'll say that flipping tails is like getting pregnant. The question is: how many of our 100 study participants will go on a 10-year "no-pregnancy" streak, by successfully flipping heads 10 times in a row?

Clearly the answer is 90%: 90 out of 100 study participants have two-headed coins. They will avoid flipping tails forever. But let's see how we could get the wrong answer by using the compounding rule instead. Suppose we proceed as follows:

1. Take data from the first year of the study, in which 90 people flip heads and 10 flip tails.
2. Calculate the average probability of successfully avoiding tails in that first year, which will be 90%.
3. Use the compounding rule to calculate the probability of a ten-year winning streak based on the one-year estimate: 0.9^{10}, or about 35%.
4. Conclude that only 35 out of 100 study participants will successfully avoid pregnancy for 10 years in a row.

This is more or less exactly what the *Times* did in its analysis of contraceptive failure rates—and it's badly wrong. Is it correct to say that the average probability of flipping heads among study participants is 0.9? Absolutely. But does that mean that the average probability of flipping heads 10 times a row is 0.9^{10}, or 35%? Absolutely not. Ten people in our study will flip tails forever, and the other 90 people will flip heads forever. The population-average probability of a 10-flip winning streak—or a

—-1
—0
—+1

streak of any length—is actually 90%, not 35%. We can't even use the compounding rule as a rough approximation. The rule just doesn't work *at all* for population averages.

Here's a second analogy, one that's much closer to our question about contraceptive effectiveness: What are the chances that you can avoid causing a car accident for the next 10 years? Every year, there are 2 million drivers in the United States who cause an accident, which is 1% of the roughly 200 million drivers in the country. Thus the "typical" American's chance of making it through a single year without causing a car accident is about 99%. To calculate the probability of making it 10 years, you might be tempted to use the compounding rule, multiplying 0.99 by itself 10 times:

$$P(\text{10-year no-crash streak}) = 0.99 \times 0.99 \times \ldots \times 0.99 = 0.904.$$

But this is wrong. To understand why, let's rewind the clock back to the end of year 1. After the first year, the American population has cleaved into two groups: 2 million people who've caused a car crash, and 198 million people who haven't. Now ask yourself two simple questions. What will happen to each group's car insurance rates, and why?

The answer is clear. Group 1, with 2 million people who caused accidents, will see their rates go up. Group 2, with 198 million people who didn't cause accidents, will see their rates stay the same or go down. Why? The insurance companies aren't doing this to punish or reward people. They're doing it to price the risk of a *future* crash appropriately—and crashes in successive years aren't independent. Past crashes predict future crashes; some people are more likely to flip heads, and others are likely to flip tails.

So what will happen in year 2? Almost surely, *more* than 1% of the people in group 1 will cause an accident in year 2. The drivers in this group are statistically less careful, at least on average. Similarly, *less* than 1% of the people in group 2 will cause an accident. The drivers in this group are statistically more careful—again, at least on average. In data science, we call this a lurking variable: something that has an important effect on the outcome of interest yet isn't directly measured.

The lurking-variable problem explains what was so wrong with our

earlier calculation, where we took the average no-crash probability of 99% and compounded it out 10 years. The question is: Whose probability were we compounding? And the answer is: Nobody's! The annual probability of 1% is a property of a population—or at best, a property of some imaginary *Homo mediocritus* who has a 1% risk of a car crash, 2.1 children, half a college degree, one testicle, and one ovary. But every *real* person has a risk that's either higher or lower than the 1% average. If you crashed in year 1, your risk looks higher; if you didn't, your risk looks lower. The compounding-rule calculation is wrong for *literally everyone*.

Back to the Pill

Let's now return to the 10-year failure rate of the pill. Using a 1-year success probability of 91%, together with the compounding rule, the *Times* arrived at a 39% probability for a 10-year "winning streak" without a pregnancy. But as we've learned, you can't just compound up a population-average probability, because doing so doesn't account for lurking variables. And there's a really important lurking variable here: some people don't use a method the way they're supposed to, so they're more likely to "flip tails" and get pregnant early in the study. Other people are consistent users, so they're more likely to "flip heads" one year after another, avoiding pregnancy through the end of the study.

In fact, there really is no such thing as a typical *user* in a contraceptive study, only a typical *group*.[14] Contraceptive research isn't like some voyeuristic game of baseball, where scientists search the nation's bedrooms high and low for the average Major League player, who hits .250. It's about waiting and counting: you follow a typical group of people who are using a method—some of them erratically, some of them consistently—and you count how many of them get pregnant over time.

In any kind of situation like this, if you use the compounding rule to reason about what will happen to the group based on their 1-year average, you will get the wrong answer. To continue with our earlier analogies:

- In the first coin flip of our hypothetical coin-flipping study, 10% of people will flip tails. That figure includes both two-headed and two-tailed coins. So if you don't flip tails, we've learned something

—-1
—0
—+1

about your coin: it has two heads. Your chance of getting tails on the next flip is 0%.

- In any given year, about 1% of Americans will cause a car accident. That figure includes both good drivers and bad drivers. So if you don't cause a crash this year, we've learned something about your driving habits. Your chance of causing a crash next year is less than 1%.
- In their first year, about 9% of typical pill users will get pregnant. That figure includes both consistent and erratic users. So if you don't get pregnant this year, we've learned something about your pill-using habits. Your chance of getting pregnant next year is probably less than 9%. Maybe it's 8%, maybe it's 2%—nobody knows, because nobody's done the study. But we do know that the most erratic users, who contributed the most toward the 9% year-1 failure rate, have now dropped out.

Figure 7.2 conveys this idea. It compares the cumulative 10-year pregnancy rates among typical pill users, under two sets of assumptions. The dotted-line curve assumes what the *Times* assumed: that women who remain in the study in later years keep getting pregnant at the same unrealistically high rate of 9% per year. This predicts a cumulative failure rate of 61% over 10 years, with failures happening just as frequently in year 10 as in year 1.

Meanwhile, the black curve assumes that in later years, the women remaining in the study have *less* than a 9% average chance of getting pregnant, since the least-adherent users have already dropped out. This effect gets stronger over time, so that by the end of the study, only the most careful users remain. This curve predicts a cumulative pregnancy rate of more like 25% over 10 years, with the vast majority of contraceptive failures happening early in the 10-year window, and to erratic users.

We should emphasize that the only thing researchers actually know from the data is that 9% of pill users in a "typical use" cohort will get pregnant in year 1. From year 2 onward, both curves are extrapolations, based only on modeling assumptions.

-1—
0—
+1—

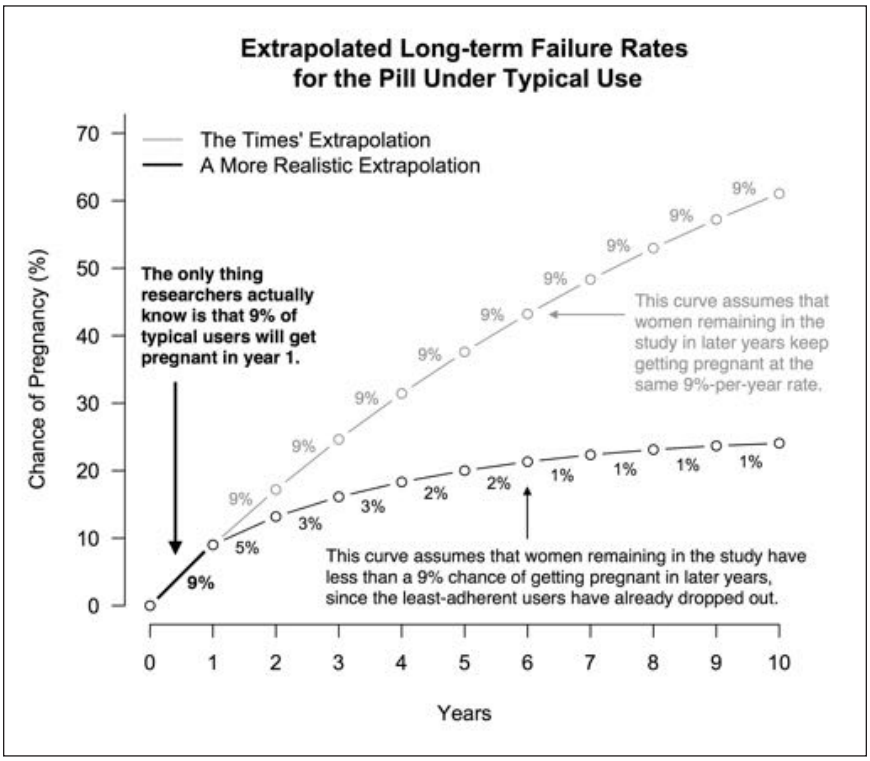

Figure 7.2

But while all models are wrong, some models are more wrong than others.

Epilogue: A "Most Disastrous and Fruitless Mania"

We see the *Times* article on contraceptive failure rates as an example of what data-analysis guru Edward Tufte once called the "rage-to-conclude bias." He took the name from an aphorism of Flaubert's: "The rage for wanting to conclude is one of the most disastrous and fruitless manias to befall humanity."[15]

Tufte was referring to the human tendency to see patterns in randomness, but the rage-to-conclude phenomenon certainly doesn't stop there. Sometimes a data set is inherently unable to answer a question.

When that happens, you really should go find data that *can* answer it. For example, data on the 1-year failure rate of the pill can't tell you about the 10-year failure rate; to learn what happens after 10 years, you need to wait 10 years. But if you're really raging to know the answer *right now,* it's regrettably tempting to torture a confession out of the data you've got, using dubious assumptions. That confession might end up doing real damage. It's one thing to use idealized assumptions to analyze a Joe DiMaggio hitting streak; not much rides on the outcome. It's another thing entirely to use those same assumptions to analyze contraceptive effectiveness—a domain where those assumptions are *very* wrong, and where fake news might harm millions of people.

This may have been a small-data mistake, but the lesson for the big-data world of AI is clear. Imagine now that those poor assumptions aren't merely used to write a one-off newspaper article. Instead, they're baked into an AI system that makes automatic decisions without a human in the loop. That's exactly how you end up with situations like the following.

- In April of 2011, there were 17 copies available on Amazon of *The Making of a Fly,* a classic book about developmental biology. The cheapest of 15 used copies was $35.54, while the cheapest of two new copies was more than $23 million. It turned out that two algorithms, run by two different book sellers, had gotten into an inverse bidding war, under poor assumptions about the behavior of other sellers.[16]
- An online clothing retailer called Solid Gold Bomb created an algorithm that automatically made new designs for print-on-demand T-shirts, based on inserting random phrases into popular slogans, such as "Keep Calm and Carry On." Because of poor oversight, the company ended up accidentally advertising T-shirts emblazoned with terrible misogynistic phrases, including ones about sexual assault. It was a traumatic experience for many who encountered the designs online, and the company went out of business because of the backlash.[17]
- On May 6, 2010, U.S. stocks experienced a "Flash Crash," in which the market lost a trillion dollars of value in a matter of minutes—