The Gift of Discovery

November 25, 2019

For many of us, December is a month of festivity and gift giving, so it's time to muse about what would make the perfect gift for a psychological scientist (other than tenure, or your paper on the cover of *Nature*). My pick would be a new scientific tool — the sort that extends what our eyes, ears, and brains can observe, creating opportunities to ask new questions. What better gift to yourself than a means to make discoveries?

Tools have a central place in the history of our science. In the 19th century, psychology was transformed into a full-fledged empirical science by importing tools from neurology and physiology. These tools helped our predecessors measure changes in heart rate, fashion reaction-time experiments, and observe behaviors in patients with brain lesions. Intrepid investigators could then search for the natural basis of mental categories that describe various types of knowing (cognition), feeling (emotions), and action (volition).

Since that heady time, psychological science has been continually enriched by periodic infusions of contraptions and computations from other disciplines. When I started my career, structural equation modeling was just coming into its own, functional MRI was peeking over the horizon, optogenetics didn't even exist, and the web was barely a glimmer on our 13-inch CRT monitors. Fast-forward 25 years, and these miracles of technology are part of our standard operating toolbox. As that toolbox expands, so do the boundaries of what we can discover.

It's worth remembering, though, that tools can obscure as easily as they can reveal. For example, our field includes a diversity of views on what a mind is and how it's implemented by the brain and body. Tools can silently amplify a scientist's deeply held beliefs as readily as they can provide unintuitive revelations.

A prime example is one of the newest additions to the psychologist's toolbox: machine learning. This

powerful family of techniques from artificial intelligence (AI) lets us make decisions and infer outcomes by training a computer model on one set of observations, identifying patterns, and generalizing to a new set of observations. The promise of machine learning (or, depending whom you ask, its hype) lies in its supposed objectivity, the possibility of removing the human mind from the equation.

Strictly speaking, machine learning is not completely new to us. Regression analysis is a simple type of machine learning for making inferences from examples. Also, machine learning has been around for years in computer science and has seeped into every domain of life. Search engines, for example, use machine learning to fill in your query faster than you can type it, because they've learned from billions of previous searches to predict with eerie accuracy what people are looking for.

In psychology, we apply machine learning to classify images, identify preferences, and search for biomarkers. Usually we assume that our models are discovering something true in the natural world that does not depend on our own thoughts and beliefs . . . but are we?

I'll show you what I mean with a toy example. Suppose we want to distinguish preferences for cake from those for cookies. (What's not to love about both, particularly if they are made with chocolate, I know, but bear with me.) In phase one of our imaginary study, we use our intuitions about cakes and cookies to collect thousands of photos of delectable baked goods from Google Images: gooey chocolate chip cookies, slices of buttery pound cake, luscious chocolate cheesecake, and other yummy delights. To be on the safe side, we ask hordes of participants to confirm our stimulus selection by explicitly labeling each photograph as "cake" or "cookie."

In phase two, we ask even more participants to rate the depicted cakes and cookies on how appetizing they are, with questions like, "How much would you enjoy eating this?" or "How tasty is this?" We crowdsource the rating task using Amazon's Mechanical Turk, and data collection is finished within a few weeks. Then we use the responses to train a supervised machine-learning algorithm to classify "love of cakes" and "love of cookies." If we are successful, we have a machine-learning system that distinguishes two different sentiments with high accuracy. That means we can apply the resulting system to classify new samples. We can generate pretty graphs showing how the data clusters beautifully into categories that reflect preferences for cakes and cookies. We can even scan participants' brains as they make their ratings, applying machine learning to find patterns of brain activity. And voilà, we have discovered the biomarkers for "love of cake" and "love of cookies."

So where's the controversy?

This style of empirical approach has many issues we could discuss, to be sure, but here we'll focus on just one: We assumed from the beginning that cookies and cakes are distinct categories. This assumption guided our stimulus section, our study design, and the data we collected. And lo and behold, the results of our supervised machine-learning analysis are consistent with the two categories we *stipulated* at the start. Our common-sense ideas became encoded in the machine-learning model during training.

What if we had sampled a wider variety of desserts — would these same two categories have emerged? Perhaps the cake/cookie distinction is more like a continuum, with (say) a classic chocolate chip cookie on one end and a devil's food cake on the other. But what about brownies — are they cookies or cakes? How about a chocolate cookie bar, or a madeleine? What about minidonuts or coconut macaroons, for

god’s sake?

Similarly, what if we had let the machine-learning algorithm infer categories instead of using the “cake” and “cookie” labels we provided (a technique called *unsupervised machine learning*)? Would we have instead discovered categories other than “cake” and “cookie,” such as “vanilla” and “butterscotch,” or a “craving for desserts with ripples” and “craving for desserts with shiny surfaces”? (Unsupervised machine learning is not really a solution on its own, however, because it can only cluster on the data that we feed it.)

Our study, by failing to consider that they might be situated in context, also assumed that preferences for cookies and cakes are immutable categories. Maybe some people prefer cookies during the day and cake at night. Perhaps some people decline dessert when they feel full, while others are ready to devour chocolate under any circumstances.

Sophisticated tools do not protect us from embedding our assumptions about the world and ourselves in our stimuli, our experimental designs, and our theoretical inferences — especially tools that may give the appearance of objectivity. This point might seem obvious for our toy example of cakes versus cookies, but how about anger versus fear? Perception versus memory? Neurological versus social pain? Vision versus audition? Male versus female? Bias in the training data is just one of many issues that can lead us to seeming discoveries about the nature of the mind and behavior that unknowingly reify our own beliefs. In psychology, this larger problem is called naive realism — the belief that the world is as it appears to you. William James, in 1890, called it “the psychologist’s fallacy.” In computer science, whole fields of study have emerged, such as ethical AI and machine-learning fairness, to reduce this sort of bias. As psychological scientists, we know to keep our biases in check, and machine learning is a new domain for us to strengthen those skills.

So this holiday season, treat yourself to a new scientific tool and pursue the thrill of discovery. After you read the operating instructions, but before you power up the tool, take a moment to remember that *what* we observe depends, in large part, on *how* we observe.

And then have a piece of cake.