INELIGIBLE TO SERVE

Getting a Job

A few years ago, a young man named Kyle Behm took a leave from his studies at Vanderbilt University. He was suffering from bipolar disorder and needed time to get treatment. A year and a half later, Kyle was healthy enough to return to his studies at a different school. Around that time, he learned from a friend about a part-time job at Kroger. It was just a minimum-wage job at a supermarket, but it seemed like a sure thing. His friend, who was leaving the job, could vouch for him. For a high-achieving student like Kyle, the application looked like a formality.

But Kyle didn't get called back for an interview. When he inquired, his friend explained to him that he had been "red-lighted" by the personality test he'd taken when he applied for the job. The test was part of an employee selection program developed by Kronos, a workforce management company based outside of Boston. When Kyle told his father, Roland, an attorney, what had happened, his father asked him what kind of questions had appeared on the test. Kyle said that they were very much like the "Five Factor Model" test, which he'd been given at the hospital. That test grades people for extraversion, agreeableness, conscientiousness, neuroticism, and openness to ideas.

At first, losing one minimum-wage job because of a questionable test didn't seem like such a big deal. Roland Behm urged his son to apply elsewhere. But Kyle came back each time with the same news. The companies he was applying to were all using the same test, and he wasn't getting offers. Roland later recalled: "Kyle said to me, 'I had an almost perfect SAT and I was at Vanderbilt a few years ago. If I can't get a part-time minimum-wage job, how broken am I?' And I said, 'I don't think you're that broken.' "

But Roland Behm was bewildered. Questions about mental health appeared to be blackballing his son from the job market. He decided to look into it and soon learned that the use of personality tests for hiring was indeed widespread among large corporations. And yet he found very few legal challenges to this practice. As he explained to me, people who apply for a job and are red-lighted rarely learn that they were rejected because of their test results. Even when they do, they're not likely to contact a lawyer.

Behm went on to send notices to seven companies—Finish Line, Home Depot, Kroger, Lowe's, PetSmart, Walgreen Co., and Yum Brands—informing them of his intent to file a class-action suit alleging that the use of the exam during the job application process was unlawful.

The suit, as I write this, is still pending. Arguments are likely to focus on whether the Kronos test can be considered a medi cal exam, the use of which in hiring is illegal under the Americans with Disabilities Act of 1990. If this turns out to be the case, the court will have to determine whether the hiring companies themselves are responsible for running afoul of the ADA, or if Kronos is.

The question for this book is how automatic systems judge us when we seek jobs and what criteria they evaluate. Already, we've seen WMDs poisoning the college admissions process, both for the rich and for the middle class. Meanwhile, WMDs in criminal justice rope in millions, the great majority of them poor, most of whom never had the chance to attend college at all. Members of each of these groups face radically different challenges. But they have something in common, too. They all ultimately need a job.

Finding work used to be largely a question of whom you knew. In fact, Kyle Behm was following the traditional route when he applied for work at Kroger. His friend had alerted him to the opening and put in a good word. For decades, that was how people got a foot in the door, whether at grocers, the docks, banks, or law firms. Candidates then usually faced an interview, where a manager would try to get a feel for them. All too often this translated into a single basic judgment: Is this person like me (or others I get along with)? The result was a lack of opportunity for job seekers without a friend inside, especially if they came from a different race, ethnic group, or religion. Women also found themselves excluded by this insider game.

Companies like Kronos brought science into corporate human resources in part to make the process fairer. Founded in the 1970s by MIT graduates, Kronos's first product was a new kind of punch clock, one equipped with a microprocessor, which added up employees' hours and reported them automatically. This may sound banal, but it was the beginning of the electronic push (now blazing along at warp speed) to track and optimize a workforce.

As Kronos grew, it developed a broad range of software tools for workforce management, including a software program, Workforce Ready HR, that promised to eliminate "the guesswork" in hiring, according to its web page: "We can help you screen, hire, and onboard candidates most likely to be productive—the best-fit employees who will perform better and stay on the job longer."

Kronos is part of a burgeoning industry. The hiring business is automating, and many of the new programs include personality tests like the one Kyle Behm took. It is now a $500 million annual business and is growing by 10 to 15 percent a year, according to Hogan Assessment Systems Inc., a testing company. Such tests now are used on 60 to 70 percent of prospective workers in the United States, up from 30 to 40 percent about five years ago, estimates Josh Bersin of the consulting firm Deloitte.

Naturally, these hiring programs can't incorporate information about how the candidate would actually perform at the company. That's in the future, and therefore unknown. So like many other Big Data programs, they settle for proxies. And as we've seen, proxies are bound to be inexact and often unfair. In fact, the Supreme Court ruled in a 1971 case, *Griggs v. Duke Power Company*, that intelligence tests for hiring were discriminatory and therefore illegal. One would think that case might have triggered some soul-searching. But instead the industry simply opted for replacements, including personality tests like one that red-flagged Kyle Behm.

Even putting aside the issues of fairness and legality, research suggests that personality tests are poor predictors of job performance. Frank Schmidt, a business professor at the University of Iowa, analyzed a century of workplace productivity data to measure the predictive value of various selection processes. Personality tests ranked low on the scale—they were only one-third as predictive as cognitive exams, and also far below reference checks. This is particularly galling because certain personality tests, research shows, can actually help employees gain insight into themselves. They can also be used for team building and for enhancing communication. After all, they create a situation in which people think explicitly about how to work together. That intention alone might end up creating a better working environment. In other words, if we define the goal as a happier worker, personality tests might end up being a useful tool.

But instead they're being used as a filter to weed out applicants. "The primary purpose of the test," said Roland Behm, "is not to find the best employee. It's to exclude as many people as possible as cheaply as possible."

You might think that personality tests would be easy to game. If you go online to take a Five Factor Personality Test, it looks like a cinch. One question asks: "Have frequent mood swings?" It would probably be smart to answer "very inaccurate." Another asks: "Get mad easily?" Again, check no. Not too many companies want to hire hotheads.

In fact, companies can get in trouble for screening out applicants on the basis of such questions. Regulators in Rhode Island found that CVS Pharmacy was illegally screening out applicants with mental illnesses when a personality test required respondents to agree or disagree to such statements as "People do a lot of things that make you angry" and "There's no use having close friends; they always let you down." More intricate questions, which are harder to game, are more likely to keep the companies out of trouble. Consequently, many of the tests used today force applicants to make difficult choices, likely leaving them with a sinking feeling of "Damned if I do, damned if I don't."

McDonald's, for example, asked prospective workers to choose which of the following best described them:

"It is difficult to be cheerful when there are many problems to take care of" *or* "Sometimes, I need a push to get started on my work."

The *Wall Street Journal* asked an industrial psychologist, Tomas Chamorro-Premuzic, to analyze thorny questions like these. The first item, Chamorro-Premuzic said, captured "individual differences in neuroticism and conscientiousness"; the second, "low ambition and drive." So the prospective worker is pleading guilty to being either high-strung or lazy.

A Kroger question was far simpler: Which adjective best describes you at work, unique or orderly?

Answering "unique," said Chamorro-Premuzic, captures "high self concept, openness and narcissism," while "orderly" expresses conscientiousness and self control.

Note that there's no option to answer "all of the above." Prospective workers must pick one option, without a clue as to how the program will interpret it. And some of the analysis will draw unflattering conclusions. If you go to a kindergarten class in much of the country, for example, you'll often hear teachers emphasize to the children that they're unique. It's an attempt to boost their self-esteem and, of course, it's true. Yet twelve years later, when that student chooses "unique" on a personality test while applying for a minimum-wage job, the program might read the answer as a red flag: Who wants a workforce peopled with narcissists?

Defenders of the tests note that they feature lots of questions and that no single answer can disqualify an applicant. Certain patterns of answers, however, can and do disqualify them. And we do not know what those patterns are. We're not told what the tests are looking for. The process is entirely opaque.

What's worse, after the model is calibrated by technical experts, it receives precious little feedback. Again, sports provide a good contrast here. Most professional basketball teams employ data geeks, who run models that analyze players by a series of metrics, including foot speed, vertical leap, free-throw percentage, and a host of other variables. When the draft comes, the Los Angeles Lakers might pass on a hotshot point guard from Duke because his assist statistics are low. Point guards have to be good passers. Yet in the following season they're dismayed to see that the rejected player goes on to win Rookie of the Year for the Utah Jazz and leads the league in assists. In such a case, the Lakers can return to their model to see what they got wrong. Maybe his college team was relying on him to score, which punished his assist numbers. Or perhaps he learned something important about passing in Utah. Whatever the case, they can work to improve their model.

Now imagine that Kyle Behm, after getting red-lighted at Kroger, goes on to land a job at McDonald's. He turns into a stellar employee. He's managing the kitchen within four months and the entire franchise a year later. Will anyone at Kroger go back to the personality test and investigate how they could have gotten it so wrong?

Not a chance, I'd say. The difference is this: Basketball teams are managing individuals, each one potentially worth millions of dollars. Their analytics engines are crucial to their competitive advantage, and they are hungry for data. Without constant feedback, their systems grow outdated and dumb. The companies hiring minimum-wage workers, by contrast, are managing herds. They slash expenses by replacing human resources professionals with machines, and those machines filter large populations into more manageable groups. Unless something goes haywire in the workforce—an outbreak of kleptomania, say, or plummeting productivity—the company has little reason to tweak the filtering model. It's doing its job—even if it misses out on potential stars.

The company may be satisfied with the status quo, but the victims of its automatic systems suffer. And as you might expect, I consider personality tests in hiring departments to be WMDs. They check all the boxes. First, they are in widespread use and have enormous impact. The Kronos exam, with all of its flaws, is scaled across much of the hiring economy. Under the previous status quo, employers no doubt had biases. But those biases varied from company to company, which might have cracked open a door somewhere for people like Kyle Behm. That's increasingly untrue. And Kyle was, in some sense, lucky. Job candidates, especially those applying for minimum-wage work, get rejected all the time and rarely find out why. It was just chance that Kyle's friend happened to hear about the reason for his rejection and told him about it. Even then, the case against the big Kronos users would likely have gone nowhere if Kyle's father hadn't been a lawyer, one with enough time and money to mount a broad legal challenge. This is rarely the case for low-level job applicants. *

Finally, consider the feedback loop that the Kronos personality test engenders. Red-lighting people with certain mental health issues prevents them from having a normal job and leading a normal life, further isolating them. This is exactly what the Americans with Disabilities Act is supposed to prevent.

▪ ▪ ▪

The majority of job applicants, thankfully, are not blackballed by automatic systems. But they still face the challenge of moving their application to the top of the pile and landing an interview. This has long been a problem for racial and ethnic minorities, as well as women.

In 2001 and 2002, before the expansion of automatic résumé readers, researchers from the University of Chicago and MIT sent out five thousand phony résumés for job openings advertised in the *Boston Globe* and the *Chicago Tribune*. The jobs ranged from clerical work to customer service and sales. Each of the résumés was modeled for race. Half featured typically white names like Emily Walsh and Brendan Baker, while the others with similar qualifications carried names like Lakisha Washington and Jamaal Jones, which would sound African American. The researchers found that the white names got 50 percent more callbacks than the black ones. But a secondary finding was perhaps even more striking. The white applicants with strong résumés got much more attention than whites with weaker ones; when it came to white applicants, it seemed, the hiring managers were paying attention. But among blacks, the stronger résumés barely made a difference. The hiring market, clearly, was still poisoned by prejudice.

The ideal way to circumvent such prejudice is to consider applicants blindly. Orchestras, which had long been dominated by men, famously started in the 1970s to hold auditions with the musician hidden behind a sheet. Connections and reputations suddenly counted for nothing. Nor did the musician's race or alma mater. The music from behind the sheet spoke for itself. Since then, the percentage of women playing in major orchestras has leapt by a factor of five—though they still make up only a quarter of the musicians.

The trouble is that few professions can engineer such an even-handed tryout for job applicants. Musicians behind the sheet can actually perform the job they're applying for, whether it's a Dvorak cello concerto or bossa nova on guitar. In other professions, employers have to hunt through résumés, looking for qualities that might predict success.

As you might expect, human resources departments rely on automatic systems to winnow down piles of résumés. In fact, some 72 percent of résumés are never seen by human eyes. Computer programs flip through them, pulling out the skills and experiences that the employer is looking for. Then they score each résumé as a match for the job opening. It's up to the people in the human resources department to decide where the cutoff is, but the more candidates they can eliminate with this first screening, the fewer human-hours they'll have to spend processing the top matches.

So job applicants must craft their résumés with that automatic reader in mind. It's important, for example, to sprinkle the résumé liberally with words the specific job opening is looking for. This could include positions (sales manager, chief financial officer, software architect), languages (Mandarin, Java), or honors (summa cum laude, Eagle Scout).

Those with the latest information learn what machines appreciate and what tangles them up. Images, for example, are useless. Most résumé scanners don't yet process them. And fancy fonts do nothing but confuse the machines, says Mona Abdel-Halim. She's the cofounder of Resunate.com, a job application tool. The safe ones, she says, are plain vanilla fonts, like Ariel and Courier. And forget about symbols such as arrows. They only confuse things, preventing the automatic systems from correctly parsing the information.

The result of these programs, much as with college admissions, is that those with the money and resources to prepare their résumés come out on top. Those who don't take these steps may never know that they're sending their résumés into a black hole. It's one more example in which the wealthy and informed get the edge and the poor are more likely to lose out.

To be fair, the résumé business has always had one sort of bias or another. In previous generations, those in the know were careful to organize the résumé items clearly and consistently, type them on a quality computer, like an IBM Selectric, and print them on paper with a high rag content. Such résumés were more likely to make it past human screeners. More times than not, handwritten résumés, or ones with smudges from mimeograph machines, ended up in the circular file. So in this sense, the unequal paths to opportunity are nothing new. They have simply returned in a new incarnation, this time to guide society's winners past electronic gatekeepers.

The unequal treatment at the hands of these gatekeepers extends far beyond résumés. Our livelihoods increasingly depend on our ability to make our case to machines. The clearest example of this is Google. For businesses, whether it's a bed-and-breakfast or an auto repair shop, success hinges on showing up on the first page of search results. Now individuals face similar challenges, whether trying to get a foot in the door of a company, to climb the ranks—or even to survive waves of layoffs. The key is to learn what the machines are looking for. But here too, in a digital universe touted to be fair, scientific, and democratic, the insiders find a way to gain a crucial edge.

• • •

In the 1970s, the admissions office at St. George's Hospital Medical School, in the South London district of Tooting, saw an opportunity. They received more than twelve applications for each of their 150 openings each year. Combing through all those applications was a lot of work, requiring multiple screeners. And since each of those screeners had different ideas and predilections, the process was somewhat capricious. Would it be possible to program a computer to sort through the applications and reduce the field to a more manageable number?

Big organizations, like the Pentagon and IBM, were already using computers for such work. But for a medical school to come up with its own automated assessment program in the late '70s, just as Apple was releasing its first personal computer, represented a bold experiment.

It turned out, however, to be an utter failure. St. George was not only precocious in its use of mathematical modeling, it seemed, but also an unwitting pioneer in WMDs.

As with so many WMDs, the problem began at the get-go, when the administrators established the model's twin objectives. The first was to boost efficiency, letting the machine handle much of the grunt work. It would automatically cull down the two thousand applications to five hundred, at which point humans would take over with a lengthy interviewing process. The second objective was fairness. The computer would remain unswayed by administrators' moods or prejudices, or by urgent entreaties from lords or cabinet ministers. In this first automatic screening, each applicant would be judged by the same criteria.

And what would those criteria be? That looked like the easy part. St. George's already had voluminous records of screenings from the previous years. The job was to teach the computerized system how to replicate the same procedures that human beings had been following. As I'm sure you can guess, these inputs were the problem. The computer learned from the humans how to discriminate, and it carried out this work with breathtaking efficiency.

In fairness to the administrators at St. George's, not all of the discrimination in the training data was overtly racist. A good number of the applications with foreign names, or from foreign addresses, came from people who clearly had not mastered the English language. Instead of considering the possibility that great doctors could learn English, which is obvious today, the tendency was simply to reject them. (After all, the school had to discard three-quarters of the applications, and that seemed like an easy place to start.)

Now, while the human beings at St. George's had long tossed out applications littered with grammatical mistakes and misspellings, the computer—illiterate itself—could hardly follow suit. But it could correlate the rejected applications of the past with birthplaces and, to a lesser degree, surnames. So people from certain places, like Africa, Pakistan, and immigrant neighborhoods of the United Kingdom, received lower overall scores and were not invited to interviews. An outsized proportion of these people were nonwhite. The human beings had also rejected female applicants, with the all-too-common justification that their careers would likely be interrupted by the duties of motherhood. The machine, naturally, did the same.

In 1988, the British government's Commission for Racial Equality found the medical school guilty of racial and gender discrimination in its admissions policy. As many as sixty of the two thousand applicants every year, according to the commission, may have been refused an interview purely because of their race, ethnicity, or gender.

The solution for the statisticians at St. George's—and for those in other industries—would be to build a digital version of a blind audition eliminating proxies such as geography, gender, race, or name to focus only on data relevant to medical education. The key is to analyze the skills each candidate brings to the school, not to judge him or her by comparison with people who seem similar. What's more, a bit of creative thinking at St. George's could have addressed the challenges facing women and foreigners. The *British Medical Journal* report accompanying the commission's judgment said as much. If language and child care issues posed problems for otherwise solid candidates, the solution was not to reject those candidates but instead to provide them with help—whether English classes or onsite day care—to pull them through.

This is a point I'll be returning to in future chapters: we've seen time and again that mathematical models can sift through data to locate people who are likely to face great challenges, whether from crime, poverty, or education. It's up to society whether to use that intelligence to reject and punish them—or to reach out to them with the resources they need. We can use the scale and efficiency that make WMDs so pernicious in order to help people. It all depends on the objective we choose.

• • •

So far in this chapter, we've been looking at models that filter out job candidates. For most companies, those WMDs are designed to cut administrative costs and to reduce the risk of bad hires (or ones that might require more training). The objective of the filters, in short, is to save money.

HR departments, of course, are also eager to save money through the hiring choices they make. One of the biggest expenses for a company is workforce turnover, commonly called churn. Replacing a worker earning $50,000 a year costs a company about $10,000, or 20 percent of that worker's yearly pay, according to the Center for American Progress. Replacing a high-level employee can cost multiples of that—as much as two years of salary.

Naturally, many hiring models attempt to calculate the likelihood that each job candidate will stick around. Evolv, Inc., now a part of Cornerstone OnDemand, helped Xerox scout out prospects for its calling center, which employs more than forty thousand people. The churn model took into account some of the metrics you might expect, including the average time people stuck around on previous jobs. But they also found some intriguing correlations. People the system classified as "creative types" tended to stay longer at the job, while those who scored high on "inquisitiveness" were more likely to set their questioning minds toward other opportunities.

But the most problematic correlation had to do with geography. Job applicants who lived farther from the job were more likely to churn. This makes sense: long commutes are a pain. But Xerox managers noticed another correlation. Many of the people suffering those long commutes were coming from poor neighborhoods. So Xerox, to its credit, removed that highly correlated churn data from its model. The company sacrificed a bit of efficiency for fairness.

While churn analysis focuses on the candidates most likely to fail, the more strategically vital job for HR departments is to locate future stars, the people whose intelligence, inventiveness, and drive can change the course of an entire enterprise. In the higher echelons of the economy, companies are on the hunt for employees who think creatively and work well in teams. So the modelers' challenge is to pinpoint, in the vast world of Big Data, the bits of information that correlate with originality and social skills.

Résumés alone certainly don't cut it. Most of the items listed there—the prestigious university, the awards, even the skills—are crude proxies for high-quality work. While there's no doubt some correlation between tech prowess and a degree from a top school, it's far from perfect. Plenty of software talent comes from elsewhere—consider the high school hackers. What's more, résumés are full of puffery and sometimes even lies. With a quick search through LinkedIn or Facebook, a system can look further afield, identifying some of a candidate's friends and colleagues. But it's still hard to turn that data into a prediction that a certain engineer might be a perfect fit for a twelve-member consultancy in Palo Alto or Fort Worth. Finding the person to fill a role like that requires a far broader sweep of data and a more ambitious model.

A pioneer in this field is Gild, a San Francisco–based start-up. Extending far beyond a prospect's alma mater or résumé, Gild sorts through millions of job sites, analyzing what it calls each person's "social data." The company develops profiles of job candidates for its customers, mostly tech companies, keeping them up to date as the candidates add new skills. Gild claims that it can even predict when a star employee is likely to change jobs and can alert its customer companies when it's the right time to make an offer. But Gild's model attempts to quantify and also *qualify* each worker's "social capital." How integral is this person to the community of fellow programmers? Do they share and contribute code? Say a Brazilian coder—Pedro, let's call him—lives in São Paulo and spends every evening from dinner to one in the morning in communion with fellow coders the world over, solving cloud-computing problems or brainstorming gaming algorithms on sites like GitHub or Stack Overflow. The model could attempt to gauge Pedro's passion (which probably gets a high score) and his level of engagement with others. It would also evaluate the skill and social importance of his contacts. Those with larger followings would count for more. If his principal online contact happened to be Google's Sergey Brin, or Palmer Luckey, founder of the virtual reality maker Oculus VR, Pedro's social score would no doubt shoot through the roof.

But models like Gild's rarely receive such explicit signals from the data. So they cast a wider net, in search of correlations to workplace stardom wherever they can find them. And with more than six million coders in their database, the company can find all kinds of patterns. Vivienne Ming, Gild's chief scientist, said in an interview with *Atlantic Monthly* that Gild had found a bevy of talent frequenting a certain Japanese manga site. If Pedro spends time at that comic-book site, of course, it doesn't predict superstardom. But it does nudge up his score.

That makes sense for Pedro. But certain workers might be doing something else offline, which even the most sophisticated algorithm couldn't infer—at least not today. They might be taking care of children, for example, or perhaps attending a book group. The fact that prospects don't spend six hours discussing manga every evening shouldn't be counted against them. And if, like most of techdom, that manga site is dominated by males and has a sexual tone, a good number of the women in the industry will probably avoid it.

Despites these issues, Gild is just one player. It doesn't have the clout of a global giant and is not positioned to set a single industry standard. Compared to some of the horrors we've seen—the predatory ads burying families in debt and the personality tests excluding people from opportunities—Gild is tame. Its category of predictive model has more to do with rewarding people than punishing them. No doubt the analysis is uneven: some potential stars are undoubtedly overlooked. But I don't think the talent miners yet rise to the level of a WMD.

Still, it's important to note that these hiring and "onboarding" models are ever-evolving. The world of data continues to expand, with each of us producing ever-growing streams of updates about our lives. All of this data will feed our potential employers, giving them insights into us.

Will those insights be tested, or simply used to justify the status quo and reinforce prejudices? When I consider the sloppy and self-serving ways that companies use data, I'm often reminded of phrenology, a pseudoscience that was briefly the rage in the nineteenth century. Phrenologists would run their fingers over the patient's skull, probing for bumps and indentations. Each one, they thought, was linked to personality traits that existed in twenty-seven regions of the brain. Usually, the conclusion of the phrenologist jibed with the observations he made. If a patient was morbidly anxious or suffering from alcoholism, the skull probe would usually find bumps and dips that correlated with that observation—which, in turn, bolstered faith in the science of phrenology.

Phrenology was a model that relied on pseudoscientific nonsense to make authoritative pronouncements, and for decades it went untested. Big Data can fall into the same trap. Models like the ones that red-lighted Kyle Behm and blackballed foreign medical students at St. George's can lock people out, even when the "science" inside them is little more than a bundle of untested assumptions.