BOOKS

# THE BEAR MINIMUM

*How test engineers decide what counts as good enough.*

BY GIDEON LEWIS-KRAUS



*The testing industry forges an unlikely union of daredevils and the safety-minded.*

A run-in with human snacks transformed Kobuk the Destroyer's mother into a "problem bear." A native of the Alaskan wilderness, the grizzly was possibly ruined by a dumpster or a cache of campsite refuse, teaching her that food was simply there for the taking—a rewarding hack for the perennial dinnertime problem. This never ends well, and it was only a matter of time before her taste for free-range chickens got her killed by a rancher. If she was the problem, the two cubs she left behind were recruited to help with the solution.

The bear-resistant container was invented to arrest the inadvertent cultivation of problem bears. A new cooler or compost bin cannot be marketed as bear-resistant until bears have been given a fair chance to break into it and fail. The gold standard is set by a consortium called the Interagency Grizzly Bear Committee, or I.G.B.C. Its original certification program was mechanical. Testers dropped heavy weights from considerable heights onto the containers and used metal fangs to make stabs at perforation. These things, needless to say, lacked the wits of the animals they were meant to represent. The cubs, Kobuk in particular, found their calling at a West Yellowstone sanctuary where the I.G.B.C. enlisted the bears as product testers. Human employees would fill new containers with fish, peanut butter, honey, and dog biscuits, snap the locks shut, and deliver them to their ursine colleagues for compliance testing. Candidates pass by withstanding precisely sixty minutes of no-holds-barred "bear contact time" without sustaining holes larger than a quarter of an inch.

The pioneering roboticist Rodney Brooks once commented that "the world is its own best model." No bear surrogate, in other words, is as good as an actual bear. Kobuk justified his tenure on the federal payroll and earned the nickname "the Destroyer" not because he was savage (although he was) but because he was canny. He handily manipulated latches designed to defeat paws. The I.G.B.C. justifies its own tenure on the federal payroll because the definition of "bear-resistant" cannot be left to bears. We need people to decide that a product has failed if it can be opened in the buzzer-beating fifty-ninth minute of a bear attack.

The bears involved are expected to behave naturally. They bite and claw and futz. The humans involved must follow rules of their own devising. They structure experiments, agree on sensible measurement criteria, and publish evaluations. The bears represent the world as it is. The people represent the world as it could be—reassuringly free of problem bears. The tension between these roles is the organizing principle of Alex Davies's "Kobuk the Destroyer, and Other Tales from the Wild, Unseen World of Test Engineering" (Norton). The book is an appealingly droll and zippy contribution to the genre of admiring case studies in professional competence. It is also an effort to vindicate human judgment once and for all. Davies believes that the deployment of bears is an essential part of the scientific process. He wants us to appreciate the same thing about ourselves.

Test engineering might sound like a dry subject, but it turns out to be a terrific vector for the delivery of high-quality anecdotes. Davies introduces the discipline as one that has always gone in for the sorts of stunts we now associate with YouTube's finest content creators. Underwriters Laboratories, known for the ubiquitous "UL" imprimatur it provides to consumer products that pass its battery of safety and reliability exams, started out with the madcap energy of a "Looney Tunes" factory. When the organization's engineers wanted to verify that a particular safe could be trusted to



ILLUSTRATION BY LISA SHEEHAN

protect valuable documents in a fire, they stuffed a sample with magazines and loose-leaf paper, covered it with kindling, doused the whole thing with half a bathtub of kerosene, and tossed in a lit match. Further trials subjected the safe to hours in an industrial furnace and a thirty-foot dive onto a pile of artfully scattered bricks.

No quest for realism has proved too extreme. An airplane's black-box flight recorder, Davies writes, is exposed to "the level of sadism required to mimic what can happen when a jet flies at near supersonic speeds into a mountain or plunges from the stratosphere into the ocean." It is fired from forty-foot cannons into chunks of aluminum honeycomb before being drowned in a pressurized tank of salt water. These tasks require an unusual combination of brute physicality and procedural rigor. They select for nerds in shirtsleeves.

In many cases, test engineers reveal a theatrical flair. In late-nineteenth-century St. Louis, the new Eads Bridge was baptized by a series of increasingly ponderous locomotives. (This method has been updated only slightly; in the nineteen-nineties, engineers checked a Michigan bridge for corrosion by parking M60 tanks on it.) In another effort to convince the public that the Eads Bridge was sturdy, its builders marched an elephant across its span. The engineers, Davies adds, were probably aware that trains are heavier than elephants. They were equally aware of the popular folklore that "elephants could sense, and would avoid, unstable ground."

Humans aren't crazy about plummeting, either. Elisha Otis made the world safe for skyscrapers when he devised a fail-safe elevator. In 1854, he wowed visitors to an industrial exhibition by riding a platform to a height of thirty feet and then asking someone to sever its hoist cable. The slack triggered a spring mechanism that locked into notches in the rails and halted his fall. Informal daredevilry would remain a counterintuitive mainstay of safety. In the aftermath of the Second World War, the Air Force sought to assess a pilot's capacity to survive ejection and other forms of radical acceleration. John Stapp, a military surgeon, strapped himself into an open rocket sled that clocked in at more than six hundred miles an hour in only five seconds. A second and a half later, he was nudged to a stop with the force of a collision into a brick wall. His face turned purple, and his vision was reduced to what he described as a "shimmering salmon colored field"; it took him about eight seconds to draw a single breath. He landed on the cover of *Time*, the era's primary recommendation engine.

Despite a seemingly inexhaustible supply of demented volunteers, it's not always feasible to perform such feats in the wild. For as long as humans have flown, humans have flown into birds. Wilbur Wright survived the first avian strike; another early pilot, who might have thought it would be funny to surprise a flock of seagulls, did not. Few people are eager to rehearse an engine failure in midair. Presumably the same goes for the birds. The "bird ingestion test" is a classic example of a situation in which testers must operate under facsimile conditions. They rig up a hangar with a specialized cannon that shoots thawed bird carcasses into the blender of a spooled-up engine. The protocol has been formalized with great care. It begins with a single four-pound chicken, moves on to a fusillade of eight members of a smaller bird species, and concludes with sixteen tiny feather balls. The risk is deemed acceptable if an engine retains at least three-quarters of its power.

The statistical reasoning is sound, even if the ingestion test's verisimilitude is a little hand-wavy. Bird populations change; migratory patterns, or what experts call "waves of biomass," are fickle. In almost half of the bird-strike incidents investigated by the Federal Aviation Administration, the turbine's charnel house is so messy that the culprit's species remains anybody's guess. (Davies, who has a wonderful ear for niche jargon, is pleased to report that bird slurry is called "snarge.") We can be fairly certain that no snarge can be traced to chickens, which do not fly. From 1990 to 2019, Davies writes, "the FAA recorded more bearded seal strikes (one) than chicken-on-plane encounters (zero)." Testers resort to chicken artillery in the absence of a robust supply chain for osprey.

The use of blameless chickens reflects a compromise between workability and similarity. Experimental results are credible only if the procedure is practical, repeatable, and standardized. We don't fly in planes because someone pulled it off once. At the same time, the results are informative only if the procedure is a serviceable approximation of the actual circumstances we expect to face. Often, some basic gimmickry will suffice. Manufacturers assess the usability of medical equipment by putting it in the hands of its intended users, sometimes in fake hospital rooms filled with real beeping, where nurses might be interrupted by some nudnik asking if they own "the blue Chevy whose alarm is going off."

Other instances are more challenging. John Stapp, of the rocket sled, had a detached view of his body—he once noted that "to the human engineer, man is a thin, flexible leather sack filled with thirteen gallons of fibrous and gelatinous material, inadequately supported by an articulated bony framework"—but most of us regard the flesh with greater sentimentality. The safety features on cars, for example, have often been evaluated with cadavers, which aren't typically at risk of further harm. Nevertheless, Davies observes, the dead "have their weaknesses." For one thing, they do not brace for impact, and they frequently suffer from preëxisting conditions. They're also expensive and tricky to source ethically. Crash-test dummies are more readily produced, and their uniformity allows for better experimental control, but they're imperfect substitutes. Manufactured in the image of the "average" male of a skinnier era, they are intended to be one-size-fits-all. They are poor proxies for women, children, and the "average" male of today.

The testing industry, Davies emphasizes, consists of trade-offs all the way down. Although some products that made it to market were needlessly dangerous—there was never a reason for the Little Lady Hot Stove, a postwar toy, to be able to reach six hundred degrees, or for contemporaneous science kits to include actual uranium—other products aren't useful if they can't hurt you. The ideal balance isn't always straightforward. In the nineteen-sixties and seventies, sixty thousand people a year reached into their lawnmower chutes to unclog them and ended up in the emergency room. In 1974, technocrats at the newly formed Consumer

Product Safety Commission campaigned on behalf of the nation's yard workers, eventually advocating for a dead man's switch on every push-mower handle. In four studies, hundreds of volunteers moved their hands from a mower handle to the discharge chute; the results clustered at around three seconds. Once engineers confirmed that an affordable mechanism could stop the blade within that interval, three seconds became the standard. Industry lobbyists thought the rule was too stringent; safety advocates thought the blade should stop faster. The C.P.S.C. made a middle-of-the-road call, weighing minor additional costs against modest safety gains, and it prevailed.

Table saws went the other way. Davies recounts the story of an amateur woodworker named Steve Gass, who invented an electrical method to cut into America's thirty thousand annual shop mishaps. (In need of a stand-in for his own meat digits, he settled on hot dogs.) The contraption worked beautifully, but, because it added hundreds of dollars to the price of the appliance, it was soundly defeated by Big Saw.

The medicine-cabinet poisoning of children was an analogous problem with an ingenious and effective solution. Just as lawnmowers and saws can't function with rubber blades, drugs can't be replaced with placebos. They could be stored in bottles that were challenging to open, but, if the bottles were too challenging, people would just leave the caps off, making things worse. The remedy appeared in the form of the Palm N Turn container, which charted a path between the Scylla of a child's ability to access pills and the Charybdis of an adult's inability to access them. Such containers are punctiliously calibrated. They must stump at least eighty per cent of a panel of fifty little kids in the course of ten minutes, even after the children are explicitly shown the winning technique. Then, within five minutes, they must submit to the will of at least ninety per cent of a panel of a hundred adults between the ages of fifty and seventy. Since the early nineteen-seventies, deaths from unintentional childhood poisonings have fallen by four-fifths.

The bulk of "Kobuk the Destroyer" is a story of unacknowledged legislators—modest people who do thankless things for good reasons. Its characters are generally engineers, scientists, and bureaucrats, often in the employ of large corporations or the federal government, who refused to take certain menaces for granted, applied themselves with honor, ingenuity, and persistence, and brokered delicate accords between competing interests. As a cheerful tour of invisible expertise, the book earns its place alongside Michael Lewis's "The Fifth Risk" or his latest, "Blockers."

Davies has an ambition beyond narrative companionability. Testing engineers, he thinks, do more than keep us safe; they tell us something crucial about the prerequisites for progress. His account is underwritten by scholars concerned, in different ways, with science and technology, among them the sociologists Trevor Pinch, John Downer, and Charles Perrow and the philosopher Ian Hacking. This tradition emphasizes that what we understand about the world cannot be separated from how we have decided to understand it, and to what end.

Good engineers recognize that their decisions affect the evidence they gather. Kobuk is clearly a workable bear. He is near to hand and executes in the clutch. But, as far as his kin in the wilderness go, is he adequately similar? Alaska might very well have grizzlies that make Kobuk look like Winnie-the-Pooh, and, as Davies writes, "the four-year-olds certifying the medicine bottle as child-resistant are not the same four-year-olds who'll try this at your home." In some domains, this issue can be mitigated by the piecemeal scrutiny of individual units. For decades, tea examiners sampled every incoming shipment of imported tea to pass federal judgment on its taste. (There was a time when it was basically an office with a single official, in a viciously productive cycle of overcaffeination.) This process was burdensome and ridiculous, but for Lapsang souchong it was at least possible. Doing the same with AirPods would be overkill, so Apple relies on statistical samples: it can tolerate a negligible fraction of defective units, provided that the defect is tinny sound rather than a tendency to blow up in your ears. Vaccines, in turn, have a different risk profile. "Safe" and "reliable" mean different things in different contexts, and this semantic coördination is part of the institutional job.

Experimentalists also accept that laboratory conditions can be not merely unrepresentative but actively misleading. When the de Havilland company sought to divine the reliability of its Comet, the first commercial jetliner, its engineers subjected sections of the fuselage to an elaborate series of mock pressurizations, to assess their strength and then their vulnerability to fatigue. The test sections survived sixteen thou-



*"Fine, you were right—'Cavorting with the Devil' is not a good band name."*

sand cycles, apparently vindicating a projected lifetime of fifteen thousand flights. In 1954, though, a Comet broke apart over the sea after only about twelve hundred flights; three months later, another, with even fewer hours aloft, did so. As it turned out, the engineers had run their tests in an inauspicious order. The extreme-pressure test partly healed tiny cracks around the rivets, delaying their growth during the fatigue testing that followed. The experiment inadvertently made its specimen more durable than the planes it was meant to represent. It wasn't an ideal situation to be winging it, but de Havilland's engineers had little choice: the relevant behavior of stressed aluminum wasn't yet understood.

We can, and do, try to game out reality in advance. But even the best models are little match for the incomprehensibly dynamic complexity of the universe. In a fascinating section about the effects of skyscrapers on street-level weather, Davies describes the construction of an "atmospheric boundary-layer wind tunnel," where researchers blow bespoke gales through scale replicas of proposed cityscapes—complete with "balconies, trees, parapets"—to predict wind patterns. These researchers would like to spare visitors to an outdoor café the unpleasantness of gusts exceeding 5.6 miles per hour. The wind tunnel is fairly accurate, Davies writes, for rectangular buildings. The atmospheric physics of curved façades, by comparison, are often too cosmically labyrinthine to be domesticated. We would need to construct the actual skyscraper and then send people out to sip their coffees, which would defeat the entire point of testing.

The folly of this sort of thing is the joke at the heart of Tom McCarthy's novel "Remainder" or Nathan Fielder's series "The Rehearsal," but in the virtual world we do it all the time. Formula 1 teams, Davies explains, use software to optimize aerodynamics and receive only one chance each year to take their prototypes for a preliminary spin on the track. They inevitably discover some residue of reality that they haven't factored into their computations and, with that lesson in mind, refine their designs before the season begins. The cars race along in a perpetual feedback loop of improvement. If the Grand Prix gets boring, or the odds too predictable, Formula 1's governors change the rules to keep the sport lively.

In all these examples, people imagine the goals, people invent the metrics, people run the trials, and people learn from their errors. The anecdotes pile up to form a brief in defense of human judgment. It's easy to understand why Davies thinks this is needful. As he puts it, "Ours is a precarious moment, scarred by a pandemic and stunned by an accelerating climate crisis, run-amok tech companies, torrential misinformation, and eroding trust in public institutions." The political environment has been resolutely hostile to those who develop, maintain, and enforce our civilization's rules. They are—in the view of both the young, inexperienced men who worked for Elon Musk's DOGE and the older, more experienced appointees who have stripped regulatory and scientific agencies of their power—plodding, blank-faced dullards with nothing better to do than slow things down. Davies properly recasts his subjects as diligent, rigorous, imaginative public servants who protect our fingers from lawnmower blades and our children from white-hot toy ovens.

That's not all. The engineers' aspiration to stage reality inside the lab further demonstrates, Davies argues, that "perfect similarity is unattainable, and all testing relies on a leap of faith." We started with vaguely bearlike weights and vaguely clawlike hooks, then replaced them with actual bears, which we, in turn, replaced with better bears—the greatest of which has lent its name to Davies's book. The whole endeavor has something noble and priestly about it: the work is never done.

"The fact that the difference between a test and reality can be minimized but never eliminated is a fundamental, underappreciated reality," Davies writes, with a little more epistemological grandiosity than one ordinarily encounters in a pop-science book. The persistence of this gap is why he believes we can't get along without test engineers. If we could reproduce the world with perfect fidelity, we would no longer require their services. We would just simulate for ourselves the sunlit uplands of the future. Because no simulation will ever be perfect, test engineers are indispensable. In Davies's account, trial and error is not just a good way to learn about the world but, in fact, the only way.

If "Kobuk the Destroyer" was motivated in part by the ascendancy of the Musk mind-set, it was also motivated by Musk's ultimate justification for DOGE: artificial intelligence. Many people in Silicon Valley take for granted the hopeless inadequacy of human judgment. Consider an ongoing debate about the F.D.A., regarded by tech libertarians as retrograde and risk-averse—a reliquary of fax machines and dot-matrix printers which mostly thwarts medical progress. The reason we continue to depend on expensive, time-consuming, often inconclusive clinical trials, they maintain, is that we lack a rigorous understanding of biological law. Once A.I. creates a perfect digital simulacrum of a cell, we can circumvent the bottlenecks of the physical world. Although many of A.I.'s most enthusiastic boosters have begun to walk back such claims in favor of a Formula 1-like feedback loop, it is still common to hear that an artificial superintelligence, as one observer put it, could infer general relativity from three frames of a falling apple.

There are all sorts of reasons to find the claim about falling apples and general relativity dubious. Davies makes the case that theorists and experimentalists need each other. Just as Formula 1's modellers learn from trial-lap results, aerospace engineers have learned from physicists; we now have good whiteboard accounts of what went wrong in de Havilland's fuselage tests. The prospects for digital clones remain an open question. What's more, it shouldn't matter whether a computer can ultimately infer relativity from apple footage. The goal of the tests Davies describes is not perfect mimicry. It is to reproduce enough of the world to serve our evolving purposes—to have pill bottles that children can't open, airplanes that don't fall apart, cafés where our newspapers don't blow away, and racecars that go zoom but not boom. The hard part is deciding what, exactly, a model needs to get right and what trade-offs among competing goods we are prepared to accept. Those decisions, we might hope, will continue to be made by neither computers nor bears but people who can be asked to answer for them. ♦

# WEB DISPLAY



Even the best models are little match for the incomprehensibly dynamic complexity of the universe.

Illustration by Lisa Sheehan

URL slug magazine/2026/10/12/kobuk-the-destroyer-alex-davies-book-review

STORY How Test Engineers Decide What Counts as Good Enough

Test engineers stage reality, sometimes by extreme means, to make sure airplanes don't fall apart, children can't open pill bottles, and bears won't break into coolers.

SITE How Test Engineers Decide What Counts as Good Enough

Test engineers stage reality, sometimes by extreme means, to make sure airplanes don't fall apart, children can't open pill bottles, and bears won't break into coolers.

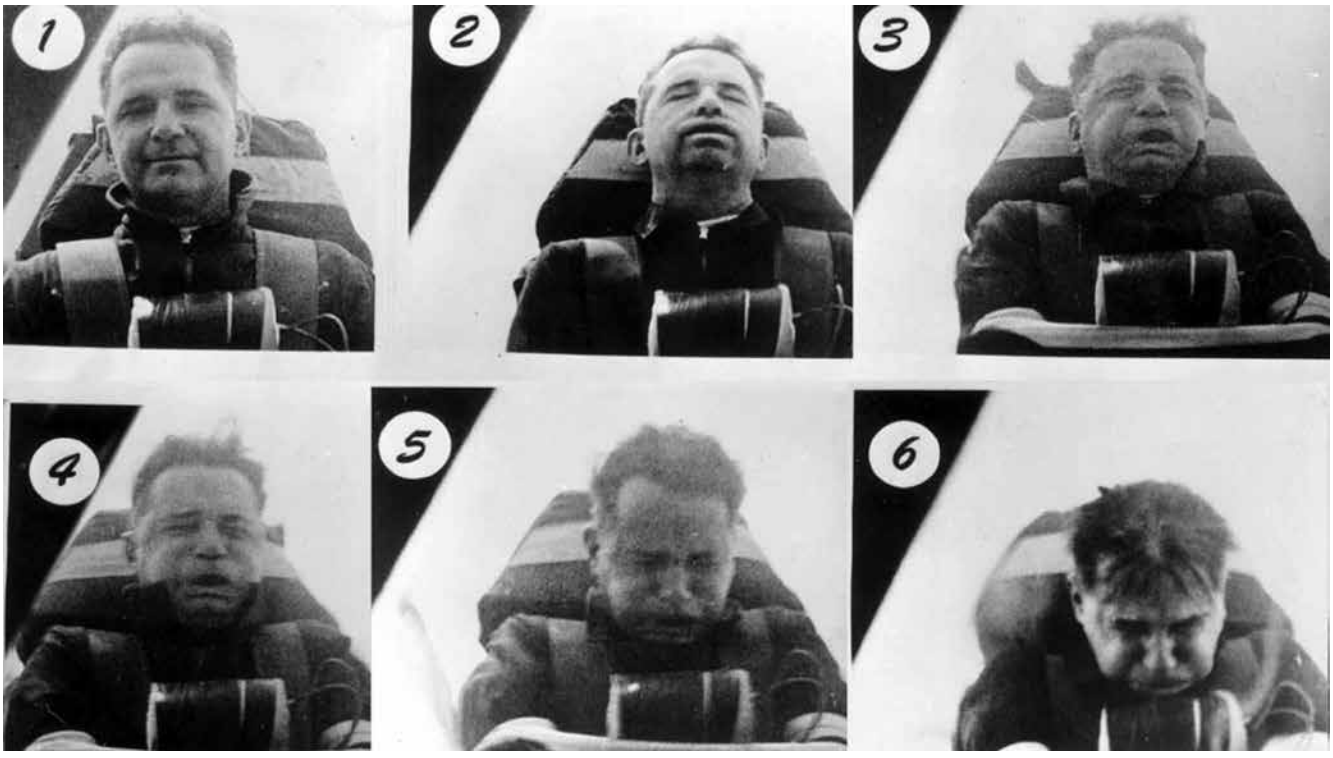
SOCIAL How Test Engineers Decide What Counts as Good Enough

Test engineers stage reality, sometimes by extreme means, to make sure airplanes don't fall apart, children can't open pill bottles, and bears won't break into coolers.

SEO How Test Engineers Decide What Counts as Good Enough

Gideon Lewis-Kraus reviews "Kobuk the Destroyer, and Other Tales from the Wild, Unseen World of Test Engineering," by Alex Davies.

inline




r50360

In March, 1954, John Stapp rode the Sonic Wind 1 rocket sled to about 418 m.p.h. after its nine solid-fuel rockets produced forty thousand pounds of thrust for five seconds. The test measured the effects of extreme acceleration on the human body.

Photograph from U.S. Air Force / Alamy