Rational Accidents

Reckoning with Catastrophic Technologies

By: John Downer

Citation:

Rational Accidents: Reckoning with Catastrophic Technologies
By: John Downer
DOI: 10.7551/mitpress/8844.001.0001
ISBN (electronic): 9780262377010
Publisher: The MIT Press
Published: 2024

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open

2 FINITISM AND FAILURE: ON THE LOGICAL IMPLAUSIBILITY OF ULTRAHIGH RELIABILITY

> Wisdom sets bounds even to knowledge.
> —Nietzsche

2.1 ON THE LIMITS OF OBJECTIVITY

REASON TO DOUBT

Japan's experience with Fukushima testifies to the importance of expert reliability assessments, and the enormous trust placed in them. Of the many engineering assertions that feed into our public policy—informing decisions and shaping priorities—few are as consequential as the reliability claims made about catastrophic technologies.

As well as testifying to the importance of these claims, however, Fukushima speaks to their fallibility. Simply put, the experts who designed and evaluated the plant were wrong in their assessment of its failure behavior. And if they were wrong with respect to Fukushima, then why not elsewhere? This question is worth examining, not least because there are good reasons to imagine it has an uncomfortable answer. There are logical and pragmatic reasons for laypeople and bureaucracies to treat most engineering assertions as incontrovertible facts. This is undeniable. Yet there are equally logical reasons to believe that engineering assertions about ultrahigh reliability are exceptional in this regard and should not be treated the same way. As we will see, decades of scholarship on the integrity and provenance of expert knowledge suggest that the predicted failure rates of catastrophic technologies are much less trustworthy than they appear. In this area, if nowhere else, it is rational to distrust the experts.

To understand why this is, it helps to visit the science and technology studies (STS) literature and some of the basic epistemology on which that literature is premised.

A FINITIST COSMOLOGY

Chapter 1 spoke of a "positivist cosmology," wherein scientists and engineers are widely thought to interrogate the world objectively and definitively to uncover facts. Collins (1985) calls this intuitive and influential notion the "canonical rational-philosophical model" of technoscientific knowledge. As he himself explains, however, it holds little sway in modern epistemology.

Beginning in the mid-twentieth century, just as the first commercial reactors came online, logical philosophers such as Wittgenstein (2001 [1953]) and Feyerabend (1975), together with philosophically minded historians such as Kuhn (1996 [1962]), began to pick apart the scientific method. By demonstrating, in different ways, the logical impossibility of an "ideal experiment" or "perfect proof," they showed how even the most rigorous and objective "facts" necessarily rest on unprovable assumptions: for example, about the representativeness of laboratories or the trustworthiness of experimenters. The work of these scholars gave way to a new orthodoxy in epistemology, wherein even the most formal and rigorous knowledge claims were understood as negotiated, interpretive, and (thus) potentially value laden.

This understanding of knowledge goes by different terms, like "relativism," "extensionalism," "finitism," and "constructivism," most of which correspond with narrow distinctions in how it is conceived and invoked. For the purposes of this text, however, I will refer to it as "finitism," a term most often associated with Bloor (1976) in this context. And I will refer to the traditional understanding of science as objective and wholly rule governed, which finitism challenges, as "positivism." Diving into the nuances of finitism here would not move the argument forward very efficiently, so let us simply note that finitists do not hold that there are *no* fundamental, ontological truths about the world, as their critics sometimes claim (e.g., Gross and Levitt 1994). Neither do they suggest that one knowledge claim is as good as any other. They simply hold that all knowledge claims—even when constrained by logic, tests, and experiments—necessarily contain fundamental ambiguities, and therefore subjective judgments: an irreducibly social component of every "fact" (e.g., Bloor 1976; Collins and Pinch 1993; Kusch 2012).

By subverting claims to perfect objectivity, finitism paved the way for social scientists to study scientific and technical knowledge itself. Inspired by the philosophers, scholars from a range of social-scientific disciplines began to explore the interpretive labor involved in forging discrete and meaningful "facts" from the unruly fabric of reality (e.g., Bloor 1976; Collins 1985; Latour 1987; Stanford 2009; Sismondo 2010). In a broad movement that gradually coalesced under the banner of STS, these scholars began to map the myriad hidden choices that go into producing knowledge, demonstrating, in a series of epistemologically conscious histories and ethnographies, that the seemingly abstruse concerns of philosophers can have tangible consequences in real-world circumstances.

The wider body of STS literature has many facets, but especially pertinent to the argument that follows is a line of research exploring the production of engineering knowledge (e.g., Pinch and Bijker 1984; MacKenzie 1990, 1996b, 2001; Bijker et al. 1989; Latour 1996). In particular, a set of studies that examine the practical uncertainties that arise when engineers draw conclusions about real-world technological performance from imperfect tests and models (e.g., Wynne 1988; Pinch 1993; MacKenzie 1996a, 1996b; Collins and Pinch 1998). This body of literature makes it abundantly clear that the engineering knowledge on which we—publics, policymakers, and bureaucracies—base technical decisions has a much messier and more complex provenance than is widely supposed. Perhaps counterintuitively, however, it rarely argues that those decisions are materially poorer for the misconception.

This apparent contradiction, wherein finitist scholarship rarely takes issue with the positivist conception of knowledge invoked by decision-makers, speaks to a longstanding tension in the discipline of STS. And understanding this tension is useful to understanding why reliability assessments of catastrophic technologies deserve to be treated differently from other engineering claims. It is to this topic, then, that we now turn.

2.2 THE PROBLEM OF RELIABILITY

UNCERTAINTY AND EFFICACY

Among the early finitists, few were as radical in their skepticism of the scientific method as the Austrian-born philosopher Paul Feyerabend. Irreverent almost to a fault, Feyerabend wrote with an unreservedness that still leads many to underestimate the rigor of his thought. In his best-known work, *Against Method* (1975), he describes university science departments as sites of indoctrination and compares them to dogmatic religious orders. After his death from an inoperable brain tumor in 1994, the headline of his obituary in the *New York Times* described him as an “anti-science philosopher.”

Yet even Feyerabend, bête noire of twentieth-century positivism, deferred to science when it counted. He might have excoriated medical science in his writings, but after his death, his wife often spoke about his “total confidence” in the doctors who had treated him, as well as his unhesitating deference to their recommendations (Horgan 2016; Feyerabend 1995). Her point was not that her late husband had abandoned his principles in his final days, but that the “anti-science” label was misleading. His writing supports this interpretation. When read closely, it is clear that Feyerabend, who had a background in theoretical physics, was a nuanced critic of the scientific method who believed strongly in the efficacy, if not the inviolability, of most technoscientific knowledge claims. He would likely have been no friend to the modern-day antivaxxer or climate change denier.

This is simply to say that finitism, even in its most radical forms, has never rejected the practical utility of technoscientific expertise. It is true that modern epistemologists, as well as the STS scholars who build on their insights, routinely argue that expert knowledge claims are less inviolable than they appear. (In the context of engineering, for example, MacKenzie [1996a] observes that the “insiders” who produce facts and artifacts tend to be more skeptical of them than the “outsiders” who use them, simply by virtue of being privy to the ambiguities of their production.) Like Feyerabend, however, these scholars rarely challenge the *efficacy* of those claims. All of them would argue that technoscientific knowledge is invariably more useful than its alternatives.

There is an undeniable tension here, with scholars often walking a fine line, but it is not an inherent contradiction. The intellectual understanding that medical knowledge is inherently imperfect is not incompatible with the conviction that penicillin cures better than prayer, or that physicians speak more authoritatively on health than faith healers. The discipline of STS is perennially engaged in complex internal debates about finitism’s bearing on the credibility of technoscientific experts (e.g., Lynch 2017; Sismondo 2017; Collins et al. 2017; Latour 2004), but nobody in these debates endorses a posttruth society where such experts are shown no deference. All freely concede, in other words, that epistemological misgivings about proof are often inconsequential in real-world circumstances.

So it is, to return to the matter at hand, that STS studies of engineering knowledge are often comfortable arguing that publics and policymakers idealize the objectivity and inviolability of that knowledge without concluding that this idealization is dysfunctional. Given that most critiques of proof are inconsequential in real-world circumstances (and this nuance is easily lost on outsiders), it might be entirely reasonable, even from a finitist perspective, to argue that laypeople and bureaucracies should approach technological questions as if they were positivists who believed engineering assertions to be objective and inviolate.

Constant (1999) articulates the logic of this position most clearly. He accepts the argument that absolute truth is an unrealizable goal in science and engineering—all engineering facts are “immutably corrigible, hypothetical and fallible,” as he puts it (355)—but he points out that technoscientific knowledge rarely has to be perfectly *true* to be *useful*. This is especially pertinent in engineering contexts, he argues, since engineers, relative to scientists, are more concerned with “application” than “discovery,” and inherently less interested with what is “true” than with what “works” (352–355). Hence, he claims, the inherent limitations of their knowledge need not diminish its practical authority. “[T]here is profound difference between reliable knowledge and unreliable stuff” (335), he writes, and “most of our stuff more or less works most of the time,” despite the misgivings of modern epistemologists (331). (For further exploration of this general argument, see, e.g., Petroski [2008] and Vincenti [1990].)

Constant makes a good point. After all, engineering *does* work, as our lived experience consistently affirms. Cell phones connect; bridges do not collapse; cars start. Our artifacts might not always be perfect, but longstanding experience suggests that expert claims about their properties deserve to be trusted in most practical circumstances. A society that refused to heed expert assertions that a bridge was too weak to stand, or a rocket too underpowered to fly, would quickly learn some salutary lessons.

This, in essence, explains why state bureaucracies, courts, and other rule-making bodies have long been inclined to accept engineering claims as incontrovertible facts. Real engineering practice might not perfectly match its positivist caricature, but that caricature has long contributed to a functional relationship between civil society and technological expertise.

When it comes to the reliability of catastrophic technologies, however, the conventional bureaucratic relationship to engineering knowledge breaks down. Because in this context, if in few others, there are compelling reasons to believe that the epistemological limits of proof pose real dilemmas with practical consequences.

To understand why, it helps to begin by considering the nature of reliability itself.

AN UNCONVENTIONAL VARIABLE

Considered closely, reliability is a surprisingly complicated property of artifacts. As engineers use the term today, it is usually understood to represent a discrete variable: measurable and quantitatively expressible in much the same way as an artifact's mass, density, or velocity. Yet it only really assumed this meaning in the twentieth century. (Some writers identify the V2 rocket as the first industrial system for which a reliability level was intentionally defined and experimentally verified [e.g., Villemeur 1991].) Before then, reliability was conventionally understood as a qualitative virtue: more akin to "fidelity" or "trustworthiness" than to "mass," "density" or "velocity" (connotations that still linger in nonengineering usage).[1] This unusual provenance is reflected in the fact that reliability now sits awkwardly among other engineering variables, differing from them in ways that have meaningful implications.

Of these differences, four in particular are worth noting here. The first is that, unlike most variables, reliability is a contextual property of artifacts: it must be defined in reference to agreed-upon (but always contestable and potentially changeable) measures and definitions of "acceptable" functionality (Johnson 2001, 250; Shrader-Frechette 1980, 33). Reduced to its barest form, for instance, engineering textbooks often express reliability as the frequency of failures over a given time or number of operations (e.g., Bazovsky 1961). To apply even this basic definition, however, it is necessary to specify several ambiguous and context-dependent terms. The choice of whether to define reliability in respect to time or operations, for example, is a vital but qualitative distinction. (Consider, for example, a system that performs ten million operations between failures but performs a million operations per second; or, conversely, a system that goes four decades between failures but operates only twice a decade. Each system would probably be considered reliable by one metric but unreliable by the other: the first would fail every ten seconds; the second every eight operations.) Compounding this difficulty, moreover, are ambiguities in the definition of "failure" itself. Determining when a system has failed requires that we first define what constitutes normal usage and proper operation, neither of which is straightforward in every context. In the early days of the Fukushima disaster, for example, prominent experts gamely claimed that the plant had not actually failed because it had never been designed to withstand a tsunami of such magnitude (see, e.g., Sir David King, quoted in Harvey 2011).

A second distinctive property of reliability is that, when used to express the future (as opposed to past) performance of a system—as is invariably the case in catastrophic technological contexts—then it is always, on some level, an expression of certainty or confidence (i.e., in that system's future failure rate). This point is important. With most variables, engineers can compensate for uncertainty in their measurements or calculations by hedging their numbers with tools such as error bars, which independently express their confidence in those measurements or calculations. Given that reliability calculations are already expressions of confidence, however, such hedging is nonsensical. It would be pointlessly baroque for engineers to assert that they are 99 percent certain that a rocket will launch reliably, but only 50 percent certain that this assertion is accurate.

A third notable feature of reliability is that it is a negative property of artifacts. This is to say that, unlike most engineering variables, it denotes an "absence" (i.e., of failure). ("Safety is no accident," as is written in granite outside the UK Civil Aviation Authority Safety Regulation Group headquarters [Macrae 2014, 16].) The distinction is significant because absences are famously difficult to demonstrate empirically (Popper 1959). To borrow a common formulation of this well-known problem, consider that demonstrating the presence of white swans in a given territory would simply require the observation of just one white swan, whereas demonstrating the absence of black swans would require the observation of all swans in that territory and the absolute knowledge that no swans had avoided detection. The latter proposition is inherently more difficult than the former, and it would become progressively more difficult as the territory expanded. So it is that demonstrating an absence of failure becomes ever more challenging as the required mean-time-to-failure grows larger (Macrae 2014, 16). Here, time is the "territory" in which accidents are not supposed to exist, and as it grows to billions of hours of operation, it rapidly outpaces the relative size of the observable territory (i.e., the operational hours) available to experts. Proving that a system can operate for ten minutes without failing is relatively trivial in most circumstances, but proving that it can operate for a million years between failures is not.

A final point worth noting—a corollary of the previous point—is that in most contexts, reliability is a backward-looking, actuarial property of systems. Unlike most engineering variables, it is a statistical measure of a system's relationship to the world. And, as such, it becomes empirically visible only when there is statistically significant, documented experience of that system operating in the world (i.e., service data).[2] This is simply to say that claims about reliability are usually grounded in the past, even though they routinely speak to a system's future performance. Reliability assessments of infantry rifles, for example, are essentially expressions of how often those rifles have failed in service, combined with basic *ceteris paribus* assumptions about the future (that the circumstances of rifle use and manufacture will remain constant, for instance).[3]

To summarize, therefore, let us say that "reliability"—as engineers usually understand it—is a "contextual expression of actuarially derived confidence in a putative absence." As we will see, however, these properties, in combination, have complex ramifications for finitist arguments about proof, and in catastrophic technological contexts specifically, they pose seemingly insurmountable dilemmas.

To understand why this is the case, it is necessary to consider the unique demands of catastrophic technologies.

RELIABILITY AND CATASTROPHIC TECHNOLOGIES

As we have seen, catastrophic technologies have two fundamental reliability requirements. The first is that they must be *ultrareliable*, demonstrating extraordinary mean-times-to-failure. It might be true that "most of our stuff more or less works most of the time," as Constant put it, but modifiers like "most" and "more or less" have no place in discourse about catastrophic technologies (1999, 331). The second is that their reliability must be *known*, and known *prior* to their operation. They demand what Wildavsky (1988, 77–79) calls an "anticipatory" model of safety. No organization would, or legally could, deploy a new reactor or jetliner unless experts were already satisfied that its design was ultrareliable.

Together, these two requirements—that systems have ultrahigh reliabilities, and that this be established prior to their operation—impose austere demands on expert reliability calculations, exposing them much more directly to the kinds of epistemological issues that concern finitists. The reasons for this become clear if we consider how these requirements interact with the fundamental properties of reliability as a variable, as previously described. Future chapters of this book will unpack and clarify these relationships and their ramifications in more detail, but their essence can be captured in the following four observations:

> 1. *The need to establish the reliability of catastrophic technologies prior to service means that experts must derive it from tests, without recourse to actuarial service data*. In this context, engineers cannot simply extrapolate failure rates from real-world performance data, as they would with most other systems; instead, they must ground their assessments in bench tests. As we will see, however, tests are always imperfect reproductions of the real world, and extrapolating from them introduces a lot of uncertainty (or a lot more than would arise from extrapolating from service data), because it forces experts to contend with questions pertaining to the relevance and representativeness of the tests themselves.
>
> *Yet . . .*
>
> 2. *The need for ultrahigh reliability means that tests alone are insufficient for establishing the performance required of catastrophic technologies*. There are various reasons for this, as we will see, but the most straightforward of them pertain to time and resources. Engineers use statistical theories to determine the minimum sample size and test duration required to establish a given level of reliability in a system. At the extreme levels of reliability required of catastrophic technologies, however, these numbers become prohibitive. Even if many tests were run in parallel and were assumed to be perfectly representative of the real world, it would still take thousands of years to statistically establish ultrahigh reliability. (It would also be a self-defeating endeavor, given that any catastrophic failures these tests evinced would, in themselves, represent the kind of disaster that experts were explicitly seeking to avoid.)
>
> *Therefore . . .*
>
> 3. *Since engineers cannot empirically establish ultrahigh levels of reliability with tests alone, they must combine tests of a system with theoretical models of its functioning.* To transcend the practical limits of what tests can claim to demonstrate, engineers combine the tests with elaborate theoretical representations of a system's functioning. By modeling the effects of redundancy in a system, for instance, engineers can theoretically employ test results to demonstrate much higher levels of reliability than would be possible from tests alone. In doing so, however, they make their reliability calculations dependent on even deeper levels of theoretical abstraction, for they must make complex judgments about the relevance and representativeness of their models, as well as that of their tests.

This increased exposure to questions of representativeness and relevance might be manageable in some circumstances, *except that . . .*

> 4. *The ultrahigh reliability required of catastrophic technologies implies a commensurately ultrahigh degree of certainty in the correctness of any judgments made in assessing them.* Recall from the previous discussion that predictive measures of reliability can be understood as expressions of confidence (i.e., that failures will not occur). It follows from this that extremely high reliability demands extremely high levels of certainty. To establish that a system is extraordinarily unlikely to fail, in other words, experts must be extraordinarily confident of the accuracy of their tests and models. (Because when looking for potential failures over billions of hours of operation, even the smallest doubt about the relevance of a model or the representativeness of a test becomes meaningful.)[4] This need for certainty has important ramifications. In contrast to Constant's portrayal of engineering as a practical discipline, it makes the task of establishing ultrahigh reliability more akin to a search for truth than a search for utility or efficacy. In this context, if nowhere else, therefore, there is little difference between "reliable knowledge" and "reliable stuff."

In light of all this, it is reasonable to assume that experts would struggle to predictively establish the failure performance of even very straightforward systems to ultrahigh levels, and catastrophic technologies are far from straightforward. Most are highly complex, consisting of many interdependent social and technical elements that interact in nonlinear ways. (As Jerome F. Lederer, the director of manned space flight safety at the National Aeronautics and Space Administration [NASA] put it in 1968: "Apollo 8 has 5,600,000 parts and 1.5 million systems, subsystems and assemblies. With 99.9 percent reliability, we could expect 5,600 defects." [Lederer 1968]) Those elements often have unusually tight design tolerances—airplane components are extremely sensitive to mass and volume, for example—which limit the scope for engineers to compensate for uncertainty with generous margins (Younossi et al. 2001). It is not unusual, moreover, for these technologies to harness uncommon materials like rare metals or innovative composites and imperfectly understood phenomena like wind turbulence or radioactivity. (Radioactivity has a poorly modeled relationship to metal fatigue, for example, especially over long periods.) And, perhaps most distinctively, many catastrophic technologies must actively negate the effects of powerful and dangerous forces, such as gravity or fission, which are inherently unforgiving of failure and require systems to do constant work simply to remain stable.

These factors exacerbate the difficulty of technological systems to the point where even relatively low levels of reliability might be seen as an achievement, let alone those required of catastrophic technologies. Witness, for instance, the extravagantly explosive history of rocketry (e.g., Swenson, Grimwood, and Alexander 1998; Longmate 1985; Clarke 2017 [1972]; Neufeld 1990) and the reliability crisis that gripped the US defense establishment during much of the Cold War (Jones-Imhotep 2000; Coppola 1984).

EPISTEMOLOGICALLY IMPLAUSIBLE

The engineering difficulty of catastrophic technologies, combined with the epistemological difficulty of predictive, ultrahigh reliability assessments, stretch the bounds of plausibility. Taken together, they strongly suggest that finitist critiques of the "perfect proof" should matter to assertions made about the reliability of catastrophic technologies, even if we agree (for the sake of argument) that those critiques are usually irrelevant to other engineering assertions. Relative to most engineering calculations, reliability assessments of catastrophic technologies are unusually dependent on simulations—tests and models, each riddled with contestable assumptions about representativeness and relevance—and, simultaneously, they are unusually dependent on the fidelity of those simulations being perfect. These relationships greatly increase the practical significance of finitist arguments about proof, especially when combined with the complexity of the systems themselves. Decades of STS scholarship suggest it should be impossible to predict the performance of a complex system over a long timeframe, to a high degree of confidence, without recourse to historical service data. The task simply requires too many subjective judgments to be made, with too much perfection, to be plausible.

Since expert understandings of *how* reliably catastrophic technologies function hinge completely on expert understandings of how those technologies function, moreover, it follows that catastrophic technologies should be unreliable for the same reason that their reliability is unknowable. Seen from a finitist perspective, their designs represent unfathomably vast networks of knowledge claims: subtle understandings of everything from the properties and behaviors of their materials in highly specific configurations and conditions, to the nature of their operating environment over long time periods, and everything in between. These knowledge claims, in turn, rest on innumerable observations, measurements, experiments, tests, and models, each ladenwith its own theories and interpretations. And any error in any step in any of these immense chains could potentially manifest as a source of failure. Insofar as it is impossible to know, with near-absolute certainty, that every knowledge claim implicit in a system is accurate, therefore, then it should be impossible to design that system to be ultrareliable.

STS scholars have traditionally striven to remain agnostic on the truth or falsity of the knowledge claims that they explore (Bloor 1976), and, like Feyerabend, they routinely defer to expert authority in practical matters. As we have seen, however, expert claims about the reliability of catastrophic technologies are functionally claims about certainty. And if there is one matter on which STS has traditionally taken a normative position, it is certainty. Indeed, the entire discipline is premised on the understanding that logic and method, even at their most rigorous, can never completely banish uncertainty.

In this context, therefore, it is reasonable, even from an STS perspective, to take a strong position on the validity of a contemporary knowledge claim. It is not only reasonable, in fact, but necessary. The assertions that experts make about the future performance of catastrophic technologies appear to be logically incompatible with an STS understanding of engineering knowledge. Everything the discipline believes about the provenance of technoscientific knowledge suggests that we should distinguish assertions made about the reliability of catastrophic technologies from other engineering assertions and doubt their plausibility. Epistemologically, it should be impossible for experts to understand a complex technological system so thoroughly, so deeply, and so perfectly as to confidently predict its failure behavior over hundreds of millions of hours of operation in a stochastic environment. By all rights, STS scholars should greet claims of ultrahigh reliability in such technologies in the same way physicists greet claims about perpetual motion.

In a world of catastrophic technologies, where the fates of major cities hinge on assertions of ultrahigh reliability, this conclusion has sociopolitical implications that are, to borrow an endearing engineering phrase, “nontrivial.” From both a theoretical and a practical standpoint, therefore, it is important that we grapple with empirical evidence that appears to prove this conclusion wrong.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Sans and Stone Serif by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Downer, John (John R.), author.
Title: Rational accidents : reckoning with catastrophic technologies / John Downer.
Description: Cambridge, Massachusetts : The MIT Press, [2023] | Series: Inside technology | Includes bibliographical references and index.
Identifiers: LCCN 2023002845 (print) | LCCN 2023002846 (ebook) | ISBN 9780262546997 (paperback) | ISBN 9780262377027 (epub) |
ISBN 9780262377010 (pdf)
Subjects: LCSH: Reliability (Engineering) | Aircraft accidents—Prevention. | Risk assessment. | Industrial accidents—Prevention.
Classification: LCC TA169 .D69 2023 (print) | LCC TA169 (ebook) | DDC 620/.00452—dc23/eng/20230202
LC record available at https://lccn.loc.gov/2023002845
LC ebook record available at https://lccn.loc.gov/2023002846