An Unverifiable World

How bounded actors act when no oracle can say they are right

Changkun Ou

Contents

Preface: A World Without Oracles

Before sailing, marrying, or founding a city, the ancient Greeks would go to Delphi and ask the oracle. The oracle mattered not because it was always accurate, but because it promised one thing: before you acted, there was somewhere that could tell you the answer. Two thousand years later, computer scientists borrowed the word. For them, an oracle is a black box: you hand it a problem you cannot solve, and it immediately returns the correct answer. The two oracles share the same fantasy: before moving, verify right and wrong.

This book is about the world after that fantasy breaks.

We almost never verify. We act, and then, sooner or later, we find out; or we never find out at all. We think "everything can be checked" is the normal state because our earliest training comes from an unusually narrow class of tasks: arithmetic, sorting a list, checking a receipt. In those tasks the answer is close at hand, so we mistake them for the shape of the whole world. But once we leave that narrow door, verification immediately becomes a luxury. You can verify seven times eight. You cannot verify, before saying "I do," that the marriage will last; before release, that a codebase has no bugs; before committing yourself, that a theory is true, a company is healthy, or a decision is right. Most consequential

actions step onto unverified ground. The oracle does not answer, and you still have to move.

The usual responses to this condition are lament or pretense. Those who lament say that if nothing can be made certain, then every judgment is mere guesswork. Those who pretend build themselves a false oracle: they enthrone a measurable number and act as if it were the unmeasurable truth. This book does neither. It asks a more interesting question: when the oracle is absent, what do capable people actually do: scientists, engineers, mathematicians, and governors?

If you pursue that question across enough fields, you run into a surprising observation. Although the sources of unverifiability differ wildly, capable responses keep converging on the same small set.

That observation gives the book its two-layer structure. Hold it in mind, because everything later hangs from it.

The first layer: the problem is heterogeneous. Beneath the sentence "I cannot check it" lie five structurally different situations. Some have no decision procedure in principle (undecidable). Some have a procedure, but its cost explodes (intractable). Some hide the relevant state from you (partially observable). Some could be verified in principle, but you lack the time, computation, or samples (budget-constrained). Some contain an opposing system actively frustrating your check (adversarial). Treating these five as the same is the most common mistake in this territory. Part I pulls them apart.

The second layer: responses converge. Whatever face the problem wears, capable actors repeatedly reach for the same few things: replace an unmeasurable target with a measurable proxy; prove a bound on a slice you can inspect; spend expensive checks where they carry the most information; introduce an external judge;

shrink the blast radius of failure; calibrate residual risk as a probability; move checking from before the fact to after it; and use multiple independent judgments to cancel single-point error. This book calls these the eight moves, and argues that they can be gathered under four more basic levers. Part II enters four concrete sites and lets those moves appear inside their local vocabularies. Part III then lifts each move out, cleans it, names it, and lays out a cross-domain table. That table is the real payload of the book. Part IV asks why these moves, and not others.

One candid reservation has to stand at the front. Is this convergence a law, in the sense that something forces every bounded actor toward these moves? Or is it only a strong empirical pattern: something we keep seeing, but cannot prove must be so? At the moment I do not have evidence that it is a law. What this book delivers is a bounded conjecture, stated plainly, together with a shared vocabulary that can connect many fields. It is not a theorem. Chapter 14 confronts this directly.

That creates an unavoidable and fitting recursion. A book about how to act under unverifiability cannot verify its own central claim. So it can only do what it describes throughout: state a calibrated belief, draw the boundaries of the claim, invite refutation, and proceed anyway. This book will practice the methods it studies. If it is right, that self-demonstration is not a defect. It is the only honest way to write it.

One image to end on; the afterword will return to it. A ship changes course in heavy fog. The captain has charts, a compass, and estimates of the current. She does not have eyes that can see through the fog. She cannot verify, before turning the rudder, whether a reef lies ahead. The fog will not lift. The oracle will not come. But sailing cannot stop for that reason. This book wants to understand not how to wait for the fog to clear, but how a good captain actually steers inside it.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

1. H. A. Simon (1969). *The Sciences of the Artificial.* MIT Press. [2][4] Simon distinguishes natural science from "the sciences of the artificial" and argues that design is a discipline for coping with complex environments under bounded rationality. His ideas of near-decomposability, hierarchy, and satisficing provide a background for this book's central stance: actors do not verify everything; under limits of computation and information, they design responses that are good enough.
2. F. H. Knight (1921). *Risk, Uncertainty and Profit.* Houghton Mifflin. [2] Knight draws the influential line between risk, where probabilities are known and measurable, and true uncertainty, where even the probability distribution is unavailable. He then attributes entrepreneurial profit to bearing the latter. The distinction is one conceptual source for this book's use of "unverifiability."
3. N. N. Taleb (2007). *The Black Swan: The Impact of the Highly Improbable.* Random House. [2][4] Taleb argues that rare, hard-to-foresee, high-impact events dominate history and markets, while conventional bell-curve statistics systematically underestimate them. His diagnosis of prediction's limits matters here: when tail events cannot be verified in advance, changing one's exposure to surprise is often more useful than pursuing precise forecasts.
4. W. C. Wimsatt (2007). *Re-Engineering Philosophy for Lim-*

ited Beings: Piecewise Approximations to Reality. Harvard University Press. [2][3][4] Wimsatt argues that limited beings cannot possess complete truth. They rely on biased but useful heuristics, robustness analysis, and piecewise approximations to reality. This is a philosophical counterpart to the chapter's main claim and especially to the role of robustness and multiple independent routes of support.

5. J. M. Keynes (1921). *A Treatise on Probability*. Macmillan. [2] Keynes understands probability as a logical relation between propositions: the rational degree of belief given evidence. He notes that many probabilities cannot be precisely numbered and may not even be comparable. His notion of the "weight" of evidence reminds us that when evidence is thin, quantified confidence may itself be unwarranted.
6. L. J. Savage (1954). *The Foundations of Statistics*. Wiley. [2] Savage gives subjective expected utility a set of axiomatic foundations: if a person's preferences satisfy certain consistency requirements, their choices can be represented as maximizing expected utility under a subjective probability. It is the baseline for later disputes about rational choice under unverifiable outcomes.
7. D. Ellsberg (1961). "Risk, Ambiguity, and the Savage Axioms." *Quarterly Journal of Economics*, 75(4), 643-669. [2] Ellsberg's urn experiments show that people systematically prefer known probabilities to unknown ones. This ambiguity aversion violates Savage's axioms and cannot be reconciled by a single subjective probability. It gives experimental force to the Knightian distinction.
8. H. A. Simon (1955). "A Behavioral Model of Rational Choice." *Quarterly Journal of Economics*, 69(1), 99-118. [2][4] Simon introduces bounded rationality and satisficing: actors constrained by cognition and information do not enumerate all options and optimize globally. They set an aspiration level and stop when they find an option that

meets it. In unverifiable settings, "good enough" is often a rational form, not a failure.

9. H. A. Simon (1947). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*. Macmillan. [2][4] Simon treats organizations as decision structures that amplify individual bounded rationality. Organizations set premises, divide responsibilities, and build routines so members can act under incomplete information. This book extends bounded rationality from individuals to institutions.
10. A. Tversky and D. Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases." *Science*, 185(4157), 1124-1131. [2] Tversky and Kahneman show that people estimate probabilities through a small set of heuristics, such as representativeness, availability, and anchoring. These shortcuts often work but also produce predictable biases. The paper is a starting point for understanding where human judgment is reliable and where it fails.
11. D. Kahneman and A. Tversky (1979). "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47(2), 263-291. [2][4] Prospect theory models real choice through a value function relative to a reference point and nonlinear weighting of probabilities. People weigh losses more heavily than equal gains and distort small and large probabilities. It is a descriptive correction to classical expected utility.
12. D. Kahneman (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux. [2][4] Kahneman synthesizes decades of research on judgment and decision-making through the dual-process frame of System 1 and System 2. The book helps build a general picture of cognitive limits and why even experts need external correction mechanisms.
13. G. Gigerenzer and D. G. Goldstein (1996). "Reasoning the Fast and Frugal Way: Models of Bounded Rationality." *Psychological Review*, 103(4), 650-669. [2][4] Gigerenzer and

Goldstein argue that simple rules using few cues and stopping early can, in real environments, match or outperform more complex statistical models. Their work counters the idea that heuristics are merely biases and asks why simplicity can be effective.

14. G. Gigerenzer, P. M. Todd, and the ABC Research Group (1999). *Simple Heuristics That Make Us Smart*. Oxford University Press. [2][4] This collection develops the adaptive toolbox program: minds carry a set of simple heuristics tuned to particular environments, and their success depends on ecological rationality. The cases show how simple rules can make robust decisions under limited information.
15. F. A. Hayek (1945). "The Use of Knowledge in Society." *American Economic Review*, 35(4), 519-530. [2][4] Hayek argues that economically relevant knowledge is dispersed, local, and tied to particular circumstances. No central planner can gather it all; prices coordinate it in decentralized form. This reveals one source of unverifiability: the relevant information is never fully held by a single actor.
16. M. Polanyi (1958). *Personal Knowledge: Towards a Post-Critical Philosophy*. Routledge and Kegan Paul. [1][3][4] Polanyi argues that all knowing contains tacit knowledge and personal commitment. Fully objective, fully formalized knowledge is an illusion. His account explains why scientific judgment cannot be replaced entirely by rules.
17. P. E. Meehl (1954). *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press. [1][4] Meehl reviews evidence showing that simple statistical or actuarial predictions often match or exceed clinical expert judgment. It is classic evidence for outsourcing judgment to checkable rules, while warning that expert confidence and expert accuracy may diverge.
18. D. A. Schon (1983). *The Reflective Practitioner: How Pro-*

fessionals Think in Action. Basic Books. [1][4] Schon describes reflection-in-action: professionals in ambiguous and unique situations do not simply apply theory. They converse with the situation, act, and reframe the problem. This captures a professional ability that cannot be verified in advance.

19. G. A. Klein (1998). *Sources of Power: How People Make Decisions.* MIT Press. [1][4] Klein's field studies of firefighters, nurses, and other experts lead to the recognition-primed decision model. Experienced people under time pressure often generate a workable action by recognizing a pattern, then mentally simulating it. This helps explain when expert intuition can be reliable.
20. P. E. Tetlock (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press. [1][4] Tetlock tracks political and economic experts over many years and finds that their average forecasts often fall short of simple extrapolation. Cognitive style explains more than credentials: pluralistic, self-questioning "foxes" outperform single-theory "hedgehogs." The book puts expert judgment under scoreable test.
21. P. E. Tetlock and D. Gardner (2015). *Superforecasting: The Art and Science of Prediction.* Crown. [1][4] This book extends Tetlock's forecasting tournament work and describes "superforecasters": people who decompose problems, assign scoreable probabilities, update frequently, and use team correction. Forecasting appears not as a gift but as a learnable practice.
22. N. N. Taleb (2001). *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets.* Texere. [2][4] Taleb argues that people regularly mistake random outcomes for skill or necessity, especially in markets, where survivors are treated as masters. The book warns that success itself does not verify a judgment when causal

structure is uncertain.

23. N. N. Taleb (2012). *Antifragile: Things That Gain from Disorder*. Random House. [4] Taleb introduces antifragility: some systems do not merely survive volatility, but benefit from it. In an unpredictable world, he argues, one should preserve optionality and cap downside risk. This directly connects to shrinking the blast radius of failure.
24. C. E. Lindblom (1959). "The Science of 'Muddling Through'." *Public Administration Review*, 19(2), 79-88. [4] Lindblom argues that real public policy is often incremental rather than globally rational: actors make limited changes near the status quo, compare as they go, and remain tied to existing means. This legitimizes cautious correction as a reasonable response to complexity.
25. K. R. Popper (1959). *The Logic of Scientific Discovery*. Hutchinson. [3] Popper develops falsificationism: scientific theories cannot be verified as true, only exposed to possible refutation. Falsifiability, not verification, marks the boundary between science and non-science. The book is one philosophical source for the present argument.
26. T. S. Kuhn (1962). *The Structure of Scientific Revolutions*. University of Chicago Press. [1][3] Kuhn argues that science alternates between normal science under a paradigm and revolutionary shifts after accumulated anomalies produce crisis. Judgment and community matter; progress is not a simple linear accumulation of verified truths.
27. W. V. Quine (1951). "Two Dogmas of Empiricism." *The Philosophical Review*, 60(1), 20-43. [3] Quine attacks the analytic-synthetic divide and the idea that each statement faces experience alone. Beliefs meet evidence as a web; any statement can be held if changes are made elsewhere. This is a classic argument for the underdetermination of theory by evidence.
28. P. Duhem (1954). *The Aim and Structure of Physical The-*

ory. Princeton University Press. [3] Duhem argues that physical experiments never test a single hypothesis in isolation. They test a whole bundle of theory and auxiliary assumptions, so a failed prediction does not uniquely identify what is wrong. This is central to understanding why science does not rest on decisive single verifications.

29. I. Hacking (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press. [3] Hacking shifts attention from representation to intervention: when we can reliably manipulate an entity to intervene in the world, we gain reason to believe in it. His experimental realism reminds us that verification is not only passive observation; it is also action.

Part I: Unverifiability

Chapter 1: The Luxury of Verification

> **Thesis:** The complete verification that can confirm in advance that something is true, correct, or safe is the exception in the lives of humans and machines, not the norm.

Seven Times Eight, and Everything Else

You can verify that seven times eight is fifty-six. You can recount it, compute it a second way, or simply recite the multiplication table; within seconds, right and wrong are nailed down.

Now switch to a few other things. Before you say "I do," verify that the marriage will last; before you hit deploy, verify that the codebase has not a single bug; before you give half your life to it, verify that the theory you believe is true; before you take the job, verify that the company is healthy. None of these can you do. Not because you have not tried hard enough, but because things of this kind simply do not offer the option of verifying in advance.

The first cornerstone of this book is that contrast: the complete verification that can confirm in advance that something is true, right, or safe is the exception in the lives of humans and machines, not the norm. We feel it ought to be the norm only because our

intuition was shaped inside a very narrow door.

The Narrow Door Where Verification Is Cheap

Which things can we actually verify? Working an arithmetic problem, sorting a string of numbers, checking the total on a receipt, judging whether a move on the board is legal. Set these side by side and they share a few hidden features: the object is closed (everything relevant is laid out in front of you), it is finite (the cases can be counted), the answer is local and immediate (it does not depend on the distant or the future), and there exists a mechanical decision procedure (follow it and you get a yes or a no).

It is precisely this narrow door that feeds our intuition that "everything can be checked." What school rewards over and over is exactly this kind of problem: one with a standard answer, gradable on the spot. So we quietly extrapolate an experience, "in the things I have practiced, right and wrong can always be sorted out," into a worldview, "the rightness or wrongness of things can always be sorted out." That extrapolation is wrong, and wrong in a systematic way. Outside the narrow door, those four features fail almost one by one.

The Illusion Breaks in Four Places

Scale. Inside the door the cases can be counted; outside, they cannot. A program with n branches may have as many as 2^n possible execution paths, and a few dozen branches are enough to make exhaustive testing impossible to finish within the lifetime of the universe. This path explosion rules "test it all" out from the start. You can verify that it is correct on the handful of inputs you thought of; you cannot verify that it is correct on all inputs. On August 1, 2012, when Knight Capital deployed

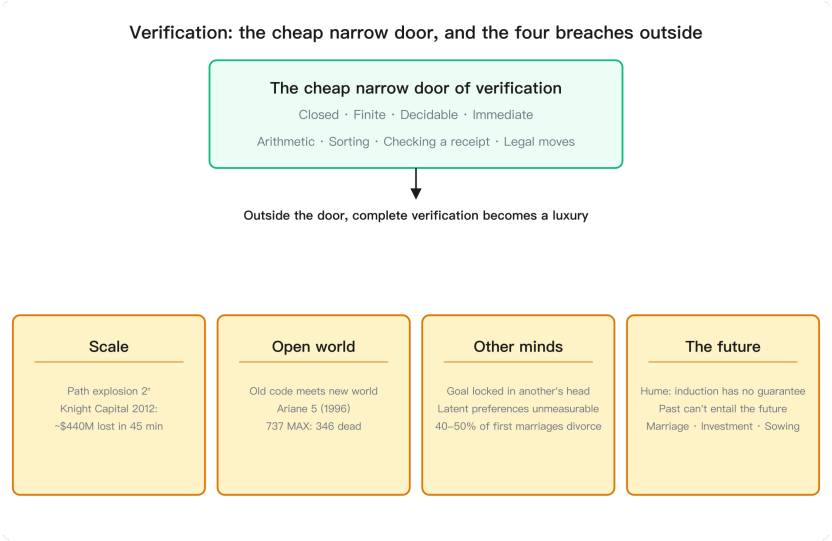

Figure 1: Verification: the cheap narrow door, and the four breaches outside it

a new trading program, one of its eight servers had not been updated, and a stretch of long-dormant, long-since-deprecated old code was accidentally woken by a reused flag; over roughly forty-five minutes after the opening bell it spewed out millions of orders, costing the firm about 440 million dollars and nearly bankrupting it overnight. No one had verified that dead path, because no one imagined it would still run. Checking a single case is easy; checking every case, the moment the quantifier "all" appears, you have crossed into another world.

Open world. Inside the door the object is closed; outside, the world keeps sending new things. What you tested was a finite handful of scenarios; what the system actually meets is an open, to-be-continued environment. On June 4, 1996, the Ariane 5 rocket made its maiden flight and destroyed itself in the air about thirty-seven seconds after launch. The cause was a piece of inertial navigation code carried over directly from the Ariane 4, never re-

verified for the new trajectory, which forced a 64-bit floating-point number into a 16-bit integer; the new rocket's higher horizontal velocity made that number overflow, and along with the four scientific satellites aboard, the loss exceeded 370 million dollars. The code had run correctly for years in the old world; in a new world it turned lethal. The MCAS system on the Boeing 737 MAX is a more harrowing version of the same breach: it behaved normally across test flight after test flight, yet on real routes it read a single faulty angle-of-attack sensor and pushed the nose down again and again; two crashes (Lion Air 610 in 2018, Ethiopian 302 in 2019) took 346 lives between them. What you verify is always the slice you have seen before; what you must wager on is the future you have not.

Other minds. Inside the door the state is observable; outside, the goal you must satisfy is often locked inside another person's head. The scene at the start of this chapter, "built it right, but it was not what I wanted," has its root here: what the user truly wants, whether the boss is satisfied, whether the other person loves you, these are latent variables; you can only infer them obliquely from behavior, you cannot read them directly, and so you cannot directly verify whether you have satisfied them. Even the most solemn of life's commitments cannot escape it: by demographic estimates, roughly forty to fifty percent of American first marriages end in divorce, and no one, in the moment of saying "I do," can verify that it will last.

The future. This is the deepest of them, and Hume laid it bare back in 1748[17]: induction carries no logical guarantee. That the sun has risen every day in the past does not logically prove it will rise tomorrow; finite past experience cannot verify in advance any universal claim about the future. What we rely on is not proof but habit. Every action whose outcome falls in the future, marriage, investment, sowing, entrusting, lies on the far side of this breach.

Even the Two Hardest Fields Bow

Perhaps you will think that the soft domains, scale and the human heart and the future, may as well concede, but mathematics and software ought to be the fortresses of complete verification. It is precisely these two hardest places that have most soberly admitted the limits of verification.

On the software side, Dijkstra[13] left a line that has been quoted to death and is still correct: testing can only prove that defects are present, never that they are absent. He held that a program should be correctly constructed, not debugged into correctness. Yet even formal proof, the strictest road, has its limit. DeMillo, Lipton, and Perlis, in their controversial and famous 1979 paper[9], argued that program verification cannot play the role mathematical proof plays, that its credibility ultimately comes from a social process rather than mechanical deduction; Fetzer in 1988 put it more harshly[10], that a program, as a causal model, is separated by a gulf from the algorithm as a logical structure, and that "a completely reliable program verification" does not hold even in theory. Brooks's "No Silver Bullet"[11] asserts that the essential complexity of software cannot be eliminated by any single stroke; Parnas resigned as an adviser to the Star Wars program and publicly argued that the software of such systems cannot be verified to the point of being worthy of trust[12]; and the Therac-25 radiotherapy machine, which between 1985 and 1987 went out of control six times because of a concurrency race condition and delivered radiation hundreds of times above normal into patients' bodies, killing at least three[15], is the footnote all these judgments paid in human lives. A 1968 NATO conference simply coined a word for it: the software crisis[16].

Mathematics cuts deeper still. Gödel proved in 1931[3] that any sufficiently rich and consistent formal system contains true propositions it cannot decide internally; Church and Turing each proved

in 1936[2,1] that no algorithm can decide whether an arbitrary proposition is provable (the Entscheidungsproblem has no solution); Rice's theorem[4] pushed it to the limit, that any nontrivial semantic property of a program is undecidable. Even where a problem is decidable in principle, the NP-completeness Cook established in 1971[5] shows that the cost of verification can explode until it simply will not run in practice. These are not temporary engineering shortfalls; they are the hard boundary logic has drawn around verification. The next chapter takes this layer apart on its own.

Restated, and This Is Not a Counsel of Despair

Put it all together: most consequential actions step onto unverified ground.

This is not a conclusion that paralyzes; it is a starting point. To admit that verification is a luxury is precisely the first step in taking action seriously. Knight, back in 1921, already separated measurable "risk" from immeasurable "uncertainty"[22] and pointed out that profit comes precisely from the latter; Keynes, speaking of genuine uncertainty, left only the line "we simply do not know"[26]; Simon, seeing that a bounded actor cannot exhaustively verify every option, proposed "satisficing"[23]; von Neumann and Morgenstern, and Savage, each built a formal framework for how to bet rationally when outcomes cannot be verified in advance[24,25]. An entire discipline of decision is built on the very premise that verification is unavailable. The question was never how to abolish uncertainty, but how to act well within it.

Where This Chapter Leads

Since verification is usually unavailable, the first thing to ask is: why is it unavailable?

There is more than one answer, and that is exactly what matters. Treating "I cannot check it" as a single situation is the most common and most misleading mistake in this field. It is in fact five structurally different situations sharing one sentence. In the next chapter, we break that sentence into five.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

1. A. M. Turing (1936). "On Computable Numbers, with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society*, s2-42, 230-265. [2] Turing, modeling computation with an abstract machine, proved that no algorithm can decide whether an arbitrary proposition is provable, and from this derived the undecidability of the halting problem. This is the founding work that raised the limit of verification from engineering experience to a mathematical theorem; the section "Even the Two Hardest Fields Bow" uses it precisely to show that the Entscheidungsproblem has no solution. The series 2, volume 42 in which it appears spans 1936 to 1937, and some bibliographies date it to 1937; the text uses the conventional 1936.

2. A. Church (1936). "An Unsolvable Problem of Elementary Number Theory." *American Journal of Mathemat-*

ics, 58(2), 345-363. [2] Church, using the lambda calculus he founded, independently proved that elementary number theory contains an unsolvable decision problem, publishing several months before Turing. His work converges with Turing's by a different route, jointly framing the theoretical boundary of "what is computable," and reminds the reader that the unavailability of verification was established along two independent paths in 1936.

3. K. Gödel (1931). "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I." *Monatshefte für Mathematik und Physik*, 38, 173-198. [2] Gödel proved that any sufficiently rich and consistent formal system contains true propositions it can neither prove nor refute. This means that "verifying every truth one by one inside the system" is impossible in principle, the deepest cornerstone of this chapter's argument that verification has a hard boundary, and one the next chapter takes apart on its own.

4. H. G. Rice (1953). "Classes of Recursively Enumerable Sets and Their Decision Problems." *Transactions of the American Mathematical Society*, 74, 358-366. [2] Rice's theorem pushes the undecidability of the halting problem to its limit: no general decision algorithm exists for any nontrivial semantic property of a program. It tells the reader that questions about "what this program will actually do" are almost uniformly not mechanically verifiable, the key support for this chapter's passage on how even the hardest fields bow.

5. S. A. Cook (1971). "The Complexity of Theorem-Proving Procedures." *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing (STOC)*, 151-158. [2] Cook here established the concept of NP-completeness, proving that the satisfiability problem carries a universal computa-

tional hardness for a large class of problems. It reveals another limit of verification: even when a problem is decidable in principle, the cost of solving or checking it may explode until it simply cannot run in practice, matching this chapter's account of how "scale" fails.

6. C. A. R. Hoare (1969). "An Axiomatic Basis for Computer Programming." *Communications of the ACM*, 12(10), 576-580. [2][1] Hoare proposed an axiomatic system, later called Hoare logic, for rigorously proving program correctness with preconditions, postconditions, and inference rules. It represents the ambition of the strictest line, to carry verification all the way through, and lets the reader see both how far formal verification can go and why, in engineering reality, it has always struggled to cover everything.

7. J. C. King (1976). "Symbolic Execution and Program Testing." *Communications of the ACM*, 19(7), 385-394. [2] King proposed symbolic execution: replacing concrete inputs with symbolic variables and systematically deriving, along a program's branches, the conditions each path must satisfy. The technique both widened the coverage of automated testing and laid bare the "path explosion" in which the number of paths grows exponentially with branches, the very difficulty this chapter uses to explain why exhaustive verification is limited.

8. E. M. Clarke and E. A. Emerson (1981). "Design and Synthesis of Synchronization Skeletons Using Branching Time Temporal Logic." *Logics of Programs (Lecture Notes in Computer Science 131)*, Springer, 52-71. [2] This workshop paper proposed using branching-time temporal logic to check automatically whether a system satisfies a given property, founding model checking. It represents the branch in which machine verification truly landed, but its power pre-

sumes a finite system state, and so it draws exactly the boundary between what automated verification can reach and what it cannot. The paper appears in LNCS volume 131, a conference proceedings rather than a journal.

9. R. A. DeMillo, R. J. Lipton, and A. J. Perlis (1979). "Social Processes and Proofs of Theorems and Programs." *Communications of the ACM*, 22(5), 271-280. [1][2] The three authors argue that mathematical proof is credible because of the social process by which the mathematical community repeatedly reads, reuses, and tests it, whereas the long and mechanical verification of programs lacks such a process and therefore cannot play the role of mathematical proof. This is the famous challenge to the idea that formal verification can give software certainty, cited here to show that the credibility of verification ultimately comes from the social rather than from pure mechanical deduction.

10. J. H. Fetzer (1988). "Program Verification: The Very Idea." *Communications of the ACM*, 31(9), 1048-1063. [1][2] Fetzer pushed the challenge deeper: an algorithm is a logical structure and can be rigorously proved, whereas a program running on a real machine is a causal model whose behavior is bound by hardware and the world, and between the two lies an unbridgeable gulf. On this basis he argued that "a completely reliable program verification" does not hold even in theory. The paper set off a large-scale debate in the 1989 Technical Correspondence, and is an important part of this chapter's demarcation of the logical boundary of verification.

11. F. P. Brooks (1987). "No Silver Bullet: Essence and Accidents of Software Engineering." *IEEE Computer*, 20(4), 10-19. [1] Brooks distinguishes the essential complexity of software from the accidental, asserting that no single technique

can yield an order-of-magnitude gain in software productivity within a decade, because essential complexity cannot be eliminated by one stroke. It supports this chapter's judgment that defects cannot be verified away in a single blow by some silver bullet. The piece was originally an invited paper for the 10th IFIP World Computer Congress in 1986, first published in Information Processing 86, 1069-1076.

12. D. L. Parnas (1985). "Software Aspects of Strategic Defense Systems." *Communications of the ACM*, 28(12), 1326-1335. [1] Parnas, after resigning as an adviser to the Star Wars program, wrote to argue point by point that the software of such systems cannot be verified, by testing or by proof, to a degree worthy of trust. This is a top engineer's public judgment on the limits of verification, made at the cost of his resignation, cited in this chapter as a real-world footnote to how even the hardest fields bow. That same year he also published a series of short essays in American Scientist.

13. E. W. Dijkstra (1972). "The Humble Programmer" (1972 ACM Turing Award Lecture). *Communications of the ACM*, 15(10), 859-866. [1] This is Dijkstra's Turing Award lecture, arguing that the programmer should stay humble, face the limited capacity of the human mind, and treat a program as something that ought to be correctly constructed rather than patched into correctness after the fact. It reflects a founder's sober judgment on the limits of after-the-fact verification, in resonance with this chapter's claim.

14. E. W. Dijkstra (1972). *Notes on Structured Programming* (in O.-J. Dahl, E. W. Dijkstra, C. A. R. Hoare, eds., *Structured Programming*). Academic Press. [1] This chapter's much-quoted yet still-correct line, "testing can only prove the presence of defects, never their absence," comes from here. Dijkstra sets out structured programming systemati-

cally, arguing that correctness is gained through disciplined construction rather than exhaustive testing. The claim first appeared in manuscript EWD249 (1970) and was formally published in *Structured Programming* in 1972.

15. N. G. Leveson and C. S. Turner (1993). "An Investigation of the Therac-25 Accidents." *IEEE Computer*, 26(7), 18-41. [1][4] The two authors give an authoritative investigation of the series of accidents in which the Therac-25 radiotherapy machine, through a software defect, overdosed patients with radiation and even killed them, analyzing the chained causes of race conditions, excessive trust in software, and the absence of an independent safety mechanism. At the cost of human lives it shows what follows when a safety-critical system is put into use without adequate verification, a heavy footnote to this chapter on the cost of verification.

16. P. Naur and B. Randell (eds.) (1969). *Software Engineering: Report on a Conference Sponsored by the NATO Science Committee*. Scientific Affairs Division, NATO. [1] This conference report records practitioners' collective anxiety that the software of the time was routinely late, over budget, and hard to deliver reliably; the term "software crisis" and the very idea of "software engineering" as a discipline came from it. It is the source of this chapter's phrase "software crisis," concentrating a generation of engineers' judgment that software could not be reliably verified. The conference was held in Garmisch, Germany, in October 1968, and the report was published in 1969.

17. D. Hume (1748). *An Enquiry Concerning Human Understanding*. (London). [4][3] Hume here laid bare the problem of induction: to infer from "it has repeatedly been so in the past" that "it will still be so in the future" carries no logical guarantee; whether the sun will rise tomorrow cannot be

proved in advance, and people act as usual out of habit, not proof. This is the source of the "future" breach in this chapter, and a starting point the whole book returns to. The 1748 first edition was originally titled *Philosophical Essays Concerning Human Understanding*, changed to the present title in 1758.

18. K. Popper (1959). *The Logic of Scientific Discovery.* Hutchinson. [3] Popper set out falsificationism systematically: a scientific theory cannot be empirically verified, only falsified, so falsifiability becomes the line between science and non-science, and science advances precisely by trying again and again to overturn theories. It bears directly on this chapter's waypoint "how science progresses," revealing that even science does not accumulate by positive verification. The English edition was expanded by the author from the German original *Logik der Forschung* (printed 1934, copyright page marked 1935).

19. W. V. O. Quine (1951). "Two Dogmas of Empiricism." *The Philosophical Review*, 60(1), 20-43. [3] Quine attacks the two dogmas of empiricism, the sharp analytic-synthetic divide and reductionism, and proposes holism: a theory is a web of belief that meets experience as a whole, and no single statement can be verified or refuted in isolation. It shows that evidence underdetermines theory, deepening this chapter's discussion of how science progresses and why verification cannot be done statement by statement.

20. T. S. Kuhn (1962). *The Structure of Scientific Revolutions.* University of Chicago Press. [3] Kuhn introduced the concept of the paradigm, describing how science alternates between the accumulation of normal science and the crisis brought on by accumulated anomalies, finally undergoing revolution through a paradigm shift. His point is

that science does not advance by the stepwise verification of truth in linear accumulation, but through incommensurable paradigm leaps. It gives this chapter's "how science progresses" a picture complementary to, and in contrast with, Popper's.

21. I. Lakatos (1976). *Proofs and Refutations: The Logic of Mathematical Discovery* (J. Worrall and E. Zahar, eds.). Cambridge University Press. [3][2] Lakatos, through a classroom dialogue tracing the evolution of Euler's formula for polyhedra, shows how mathematical concepts and theorems grow in the back-and-forth of counterexample, re-proof, and revised definition. It overturns the impression that a mathematical proof is a once-and-for-all verification, suggesting that even the most certain field advances through criticism, echoing this chapter's general account of the limits of verification.

22. F. H. Knight (1921). *Risk, Uncertainty and Profit.* Houghton Mifflin. [4] Knight separates "risk," measurable by probability, from genuine "uncertainty," which cannot be measured at all, and argues that the entrepreneur's profit comes precisely from bearing the latter. This distinction is the key to this chapter's turn, after admitting that verification is a luxury, toward a theory of action; it explains how decision and reward acquire meaning when outcomes cannot be verified in advance.

23. H. A. Simon (1955). "A Behavioral Model of Rational Choice." *The Quarterly Journal of Economics*, 69(1), 99-118. [4] Simon proposed a behavioral model of bounded rationality: an actor limited in both information and computational power cannot exhaustively compare all options, and can only set an adequate level, stopping at the first option that meets it, which is to satisfice. This is

one concrete answer to this chapter's question of how to act well within uncertainty, turning the unavailability of verification into an operable decision rule.

24. J. von Neumann and O. Morgenstern (1944). *Theory of Games and Economic Behavior*. Princeton University Press. [4] The two authors founded game theory and derived expected utility from a set of axioms, arguing that a rational actor should choose according to expected utility. It built a formal framework for how to bet rationally when the opponent's intent and the outcome cannot be verified in advance, one of the pillars of the discipline of decision, built on the unavailability of verification, that this chapter describes.

25. L. J. Savage (1954). *The Foundations of Statistics*. John Wiley & Sons. [4] Savage built axiomatic foundations for subjective probability and personalist decision theory, proving that as long as preferences satisfy certain consistency conditions, an actor behaves as if maximizing expected utility under some subjective probability and utility. It gives a standard of rationality for betting consistently in a world where probabilities cannot be objectively verified, and together with von Neumann's framework it supports this chapter's discussion of decision under uncertainty.

26. J. M. Keynes (1937). "The General Theory of Employment." *The Quarterly Journal of Economics*, 51(2), 209-223. [4] Keynes, replying to critics of the *General Theory*, stressed that genuine uncertainty cannot be measured by probability, that of some things "we simply do not know." He pointed out that an investment decision, facing an unverifiable future, can only rely on convention and animal spirits. This chapter's classic statement of genuine uncertainty comes from here, and it corroborates that the discipline of decision takes the unavailability of verification as its premise.

27. A. Tversky and D. Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases." *Science*, 185(4157), 1124-1131. [4] Tversky and Kahneman showed experimentally that in judging probability people often rely on heuristic shortcuts such as representativeness, availability, and anchoring, and so deviate systematically from the norms of probability. It completes this chapter's picture at the descriptive level: in a world where probability cannot be fully verified, how people actually judge, and where they consistently go wrong.

28. N. N. Taleb (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House. [4] Taleb calls those rare, extreme-impact events that are explained away only in hindsight "black swans," arguing that they cannot be verified or predicted in advance yet often dominate the course of history. On this basis he proposes that one give up the fantasy of precise prediction and instead build arrangements that are not destroyed by surprise, and may even benefit from it. This echoes the ending of this chapter: the problem is not to abolish uncertainty, but how to live steadily within it.

Chapter 2: The Five Faces of Unverifiability

> **Thesis:** "I cannot check it" conceals five structurally different situations, and to run them together is the central error of this field.

A single "I cannot check it" sounds like one situation, but it hides five. They are structurally different, and the remedies available to each are different too; running them together is the most central error in this field.

What this chapter sets out to do is to pry the five apart and pin each one down precisely. This may look like nothing more than a taxonomist's fastidiousness, but it is in fact the credit line for the whole second half of the book. The book ultimately wants to argue that, however wildly the sources of unverifiability differ, the responses converge on the same small set. For that claim not to look cheap, the precondition is to make "wildly different" stick first. The more thoroughly the differences are drawn, the more the later convergence becomes something worth marveling at, worth explaining. Hold these five faces firmly in mind; they will be named again and again across the book.

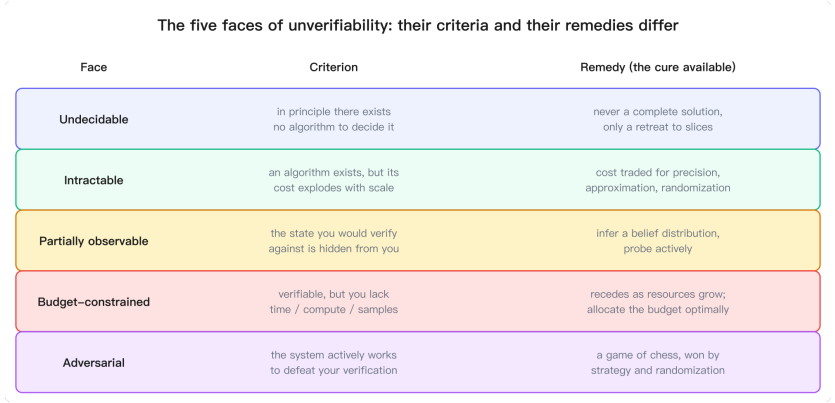

Figure 2: The five faces of unverifiability: their criteria and their remedies differ

The First Face: Undecidable

The criterion: in principle there exists no algorithm to decide it. Not hard, but absent.

This is verification's hardest failure. Hilbert and Ackermann in 1928[4] stated the decision problem (Entscheidungsproblem) explicitly, asking whether there could be a mechanical procedure to decide the truth or falsity of any mathematical proposition. Eight years later, Church[2] with the lambda calculus and Turing[1] with that abstract machine of his each proved that there could not. Turing's halting problem is especially clean: no algorithm can decide, for an arbitrary "program plus input," whether it will halt. Gödel's incompleteness of 1931[3], Rice's theorem[6] (no nontrivial semantic property of a program is decidable), and Matiyasevich's 1970 refutation[7] of Hilbert's tenth problem all belong to this family. This is no idle talk on paper: precisely because "will this stretch of program do something harmful" reduces to the halting problem, there can in theory be no antivirus software that perfectly and infallibly detects every malicious program. This is the ceiling logic has drawn for the antivirus industry.

The remedy for this face has a property unlike any other: there will never be a complete solution. No amount of time, no faster machine will do, because the obstacle is logical, not one of resources. All you can do is settle for less, verifying a finite slice, or confining yourself to those fragments that genuinely are decidable (arithmetic with addition only, say). This point will go on echoing all the way to Chapter 7.

The Second Face: Intractable

The criterion: an algorithm exists, but its cost explodes with scale, too large to finish in practice.

This face and the last are a hair apart and worlds apart: decidable, yet infeasible. The NP-completeness established independently by Cook in 1971[8] and Levin in 1973[9], and Karp's famous twenty-one NP-complete problems of 1972[10], give it a precise characterization. The satisfiability problem in the worst case, and countless combinatorial optimization problems, all have solution methods in principle, but in the worst case the time those methods take grows exponentially with the size of the input,

$$T(n) \sim 2^n,$$

and a few dozen variables are enough to leave the fastest supercomputer sighing at an ocean it cannot cross. An intuition for the order of magnitude: the game tree of chess has about 10^{120} branches (the Shannon number), the legal positions of Go number about 2×10^{170}, while the atoms in the entire observable universe come to only about 10^{80}. These board games have simple rules and are exhaustible in principle, but that "in principle" lies far beyond physical possibility. Whether P equals NP is precisely the question of whether this wall is destined to stand.

Its remedy is wholly unlike that for the undecidable. Here, putting in more resources is meaningful, and, more to the point, you can trade "accepting a little less" for "a cost you can afford": approximate solutions in place of exact ones, the average case in place of the worst case, heuristic search, randomization. Intractability forces out a whole repertoire of "discounting" wisdom, which the undecidable case has none of.

The Third Face: Partially Observable

The criterion: the very state you would verify against is hidden from you.

It is not that there is no decision procedure, nor that the cost is too great, but that you simply cannot see the thing you ought to see. The user's true preferences, what is happening inside the patient's body, the cards in the opponent's hand, these are the states that drive the outcome, yet they do not reveal themselves to you. Control theory formalized this early: Åström in 1965[15] studied optimal control under incomplete state information, and Smallwood and Sondik in 1973[16], along with Kaelbling and colleagues in 1998[18], developed it into the standard framework of the partially observable Markov decision process (POMDP). Papadimitriou and Tsitsiklis in 1987[17] further proved that solving this class of problems is itself intractable, so the third face often stacks atop the second.

Its remedy is a class of its own: you no longer pursue a definite verdict, but maintain a belief distribution over the hidden state (a belief state), and use each observation to update it,

$$b'(s') \;\propto\; \Pr(o \mid s') \sum_{s} \Pr(s' \mid s, a)\, b(s).$$

Inference and probing, rather than "computing harder," are the

cure for this face. The whole of Chapter 5 lives inside it.

The Fourth Face: Budget-Constrained

The criterion: in principle it can be verified and solved, but you, this actor, here and now, do not have the time, the computation, or the samples.

This face is the plainest and also the most common. A reviewer has only twenty minutes to read a paper; a doctor only a few minutes to make a diagnosis; a trader must place the order before the quote disappears. Verification is wholly feasible in theory, yet infeasible once it lands on a bounded actor. The bounded rationality of Knight in 1921[19] and Simon in 1955[20] is its intellectual source; the anytime algorithm of Dean and Boddy in 1988[22] (interruptible at any moment, yielding the current best solution) and the "bounded optimality" of Russell and Subramanian in 1995[23] are its formalization.

Its remedy has a feature no other face shares: this face recedes as resources grow. Give it enough time and computation and it disappears. Just for that reason, the heart of dealing with it lies in allocation, spending the scarce budget where the marginal return is highest. This line of thought is exactly where the later move of "optimal screening" comes from.

The Fifth Face: Adversarial

The criterion: the system you face is actively working to defeat your verification.

In the first four faces, the difficulty comes from a neutral property of the world: logical, of scale, of visibility, of resources. The fifth is different: across from you sits an intelligence optimizing against your check. The opponent who lies, the malicious code that dis-

guises itself, the examinee who manipulates the metric. The game theory of von Neumann and Morgenstern in 1944[24], the equilibrium of Nash in 1950 (the Nash equilibrium)[25], and the minimax criterion of Wald in 1945[26] are its classical theory; the adversarial examples discovered by Szegedy and colleagues in 2014[29], and the robust optimization with which Madry and colleagues in 2018[31] unified attack and defense, are its contemporary incarnation in machine learning. A model with an extremely high recognition rate can be fooled disastrously by a perturbation imperceptible to the human eye, because someone has gone looking specifically for that perturbation. Researchers have produced a demonstration reproduced again and again: stick a few carefully designed stickers, looking like idle graffiti, onto a stop sign, and a top-tier image-recognition system can be made to read it steadily as a speed-limit sign. The same sign: a human sees stop, the machine sees go, and the difference lies only in where the adversary places the stickers.

Its remedy is strategy, not computation. What you must do is not to compute some quantity more accurately, but to

$$\min_{x} \max_{y} L(x, y),$$

defend against the worst case, use randomization to deprive the adversary of any prediction about you, and seek to stand firm in the game rather than be optimal on some fixed input. To treat the adversarial as a mere observability gap ("I just have not seen it clearly yet") is a misjudgment that can cost lives, because it will adjust itself to track your judgment.

Five Faces, Five Cures

Set them side by side, and what matters is not the names but that their remedies are mutually incommensurable:

- The undecidable: never a complete solution, only a retreat to slices.
- The intractable: cost can be traded for precision, and more resources are meaningful.
- The partially observable: rely on inferring beliefs and active probing.
- The budget-constrained: recedes as resources grow, the heart of it lies in allocation.
- The adversarial: it is a game of chess, won by strategy and randomization.

Whoever tells you "this thing is solved by piling on more computation" has most likely mistaken one face for another. To treat the undecidable as a budget problem, or the adversarial as an observability problem, is this kind of misidentification, and a costly one.

These faces also stack and compound. The agent let loose in Chapter 6 runs into both the unpredictability of an open world (behavior that is nearly undecidable) and the strategic cunning of an opponent (adversarial); the organization in Chapter 8 bears the partially observable and the adversarial together. Real situations are often a blend of several faces.

Precisely because the sources are so uneven and the remedies so various, the next question to ask grows sharp: do humans have a mature method for coexisting with the unverifiable over the long run, with discipline? They do. That method is called science, and its very first house rule is to admit openly that it can never verify itself.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

Undecidable [2][3]

1. A. M. Turing (1936). "On Computable Numbers, with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society*, s2-42(1), 230-265. [2][3] Turing here introduces the notions of "computable numbers" and the abstract computing machine, and from the unsolvability of the halting problem derives that the decision problem has no mechanical solution. This paper is the cleanest exemplar of the "undecidable" face: the obstacle is logical, not one of resources, and this chapter takes Turing's machine and the halting problem as the standard illustration of the face.
2. A. Church (1936). "An Unsolvable Problem of Elementary Number Theory." *American Journal of Mathematics*, 58(2), 345-363. [2][3] Church, using the lambda calculus he developed, proved that elementary number theory contains an unsolvable problem and thereby independently refuted the decision problem, publishing about seven months ahead of Turing. His result and Turing's confirm each other, jointly making it stick that "in principle there exists no decision algorithm" is no isolated phenomenon; this chapter places the two side by side as the opening of the undecidable family.
3. K. Gödel (1931). "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I." *Monatshefte für Mathematik und Physik*, 38, 173-198. [2][3] Gödel here proves the incompleteness theorem: any sufficiently

strong consistent formal system contains propositions that can be neither proved nor refuted. It is the headwater of the undecidable lineage, showing that formal methods themselves have an in-principle limit, and this chapter lists it as the earliest warning bell of this face.

4. D. Hilbert & W. Ackermann (1928). *Grundzüge der theoretischen Logik*. Springer. [2][3] This textbook of mathematical logic states the decision problem explicitly for the first time, that is, it asks whether there exists a mechanical procedure that can decide the truth or falsity of any mathematical proposition. It was precisely this question that gave rise to the negative proofs of Church and Turing, and this chapter takes it as the point of departure for the undecidable face, by which the reader can see clearly the optimistic expectation of that era and the logical wall it then ran into.
5. E. L. Post (1944). "Recursively Enumerable Sets of Positive Integers and Their Decision Problems." *Bulletin of the American Mathematical Society*, 50(5), 284-316. [2][3] Post here systematically studies recursively enumerable sets and their decision problems, and proposes the line of thought that would later give rise to the theory of degrees of unsolvability. It advances the "undecidable" from a single problem to a graded study of the structure of unsolvability, and this chapter cites it to show that this face has its own levels and lineage, rather than being a monolithic block.
6. H. G. Rice (1953). "Classes of Recursively Enumerable Sets and Their Decision Problems." *Transactions of the American Mathematical Society*, 74(2), 358-366. [2] Rice's theorem establishes here that any nontrivial semantic property of what a program computes is undecidable. It generalizes Turing-style undecidability from a particular problem into a universal iron law, and this chapter cites it to show that the wish to verify mechanically whether a program "does the right thing" is, in principle, blocked off.

7. Y. V. Matiyasevich (1970). "Enumerable Sets Are Diophantine." *Soviet Mathematics. Doklady*, 11(2), 354-357. [2][3] Matiyasevich here supplies the last link, proving that every recursively enumerable set is Diophantine and thereby completing the proof of the unsolvability of Hilbert's tenth problem, that is, the MRDP theorem. It shows that even a concrete mathematical question like "whether a Diophantine equation has integer solutions" has no decision algorithm, and this chapter cites it to corroborate that the undecidable is not confined to self-reference or metamathematics but has seeped into ordinary mathematics.

Intractable [2][3]

8. S. A. Cook (1971). "The Complexity of Theorem-Proving Procedures." *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing (STOC)*, 151-158. [2][3] Cook here founds the concept of NP-completeness, proving that the satisfiability problem is the hardest class within NP. It gives a precise definition to the "intractable" face: the problem has a solution method, but the cost explodes with scale. This chapter uses it to draw the line between the second face and the first, that is, "decidable yet infeasible" differs from "no algorithm at all."
9. L. A. Levin (1973). "Universal Sequential Search Problems." *Problems of Information Transmission*, 9(3), 265-266. [2][3] Levin, on the other side of the Iron Curtain, independently obtained the same result as Cook, giving a completeness characterization of universal search problems; together the two are called the Cook-Levin theorem. It shows that the discovery of NP-completeness was a convergence rather than an accident, and this chapter cites it to reinforce the objectivity of the "intractable" face: this is a property of the structure of the problem itself, not something that varies

with the path of research.

10. R. M. Karp (1972). "Reducibility Among Combinatorial Problems." In R. E. Miller & J. W. Thatcher (Eds.), *Complexity of Computer Computations* (pp. 85-103). Plenum Press. [2][3] Karp used polynomial reduction to prove that twenty-one common combinatorial problems are all NP-complete, extending Cook's single result into a web of mutual reductions. It shows that intractability is not the quirk of a few individual hard problems but a universal phenomenon spanning a great many practical problems of scheduling, partition, covering, and the like, and this chapter cites it to show that this face is everywhere in engineering.
11. J. Hartmanis & R. E. Stearns (1965). "On the Computational Complexity of Algorithms." *Transactions of the American Mathematical Society*, 117, 285-306. [2] This paper grades algorithms by the running time of a Turing machine, laying the framework for dividing complexity classes according to resource consumption; the very term "computational complexity" was established by it. It provides the measuring stick necessary to gauge the "intractable," and this chapter cites it to show that the second face can be spoken of precisely only because there is first a language for characterizing how cost grows with scale.
12. M. R. Garey & D. S. Johnson (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman. [2] This book systematically organizes the theory of NP-completeness and its proof techniques, and appends a widely cited list of intractable problems; it has long served as the standard reference for the field. For the reader who wishes to understand the "intractable" face from the ground up, it is both an introductory guide and a working handbook, and this chapter lists it as the most reliable foothold on the subject.

13. M. Sipser (2012). *Introduction to the Theory of Computation* (3rd ed.). Cengage Learning. [2] This widely adopted undergraduate textbook clearly explains automata, computability, and complexity, threading the Turing machine, the halting problem, P and NP, and other concepts into one coherent line. It covers exactly the theoretical groundwork of this chapter's first two faces, and is the surest starting point for the reader who wishes to build a foundation from scratch; the first edition can be traced back to 1997.
14. S. Arora & B. Barak (2009). *Computational Complexity: A Modern Approach*. Cambridge University Press. [2] This graduate textbook covers everything from the classical complexity classes to modern topics such as randomization, interactive proofs, approximation, and derandomization, with a scope far beyond an introductory text. For the reader who wishes to go deep into the "intractable" face, and especially to understand how people use approximation and randomness to skirt the worst case, it is the more advanced authoritative reading.

Partially Observable [2][4]

15. K. J. Åström (1965). "Optimal Control of Markov Processes with Incomplete State Information." *Journal of Mathematical Analysis and Applications*, 10, 174-205. [2] Åström here studies optimal control under incomplete state information, proposing that all available information be summarized by a probability distribution over the hidden state, that is, a "belief state." This is one of the headwaters of POMDP theory, and exactly the cure this chapter prescribes for "partially observable": not to pursue a definite verdict, but to maintain and update a belief.
16. R. D. Smallwood & E. J. Sondik (1973). "The Optimal Control of Partially Observable Markov Processes over a

Finite Horizon." *Operations Research*, 21(5), 1071-1088. [2] This paper gives the classic structural results for the finite-horizon POMDP, and on that basis designs a method for computing the optimal policy. It advances Åström's belief-state idea into an operable algorithm, and this chapter cites it to show that "relying on inferring beliefs" is no empty phrase but is backed by established solution techniques.

17. C. H. Papadimitriou & J. N. Tsitsiklis (1987). "The Complexity of Markov Decision Processes." *Mathematics of Operations Research*, 12(3), 441-450. [2] This paper systematically characterizes the computational complexity of the variants of the Markov decision process, proving that introducing partial observability makes solving them markedly harder. It clasps the third face together with the second: the situation of not being able to see the correct state is itself often intractable to solve, and this chapter uses it precisely to show that the faces stack atop one another.
18. L. P. Kaelbling, M. L. Littman & A. R. Cassandra (1998). "Planning and Acting in Partially Observable Stochastic Domains." *Artificial Intelligence*, 101(1), 99-134. [2][4] This paper organizes the POMDP into a standard framework within artificial intelligence, unifying belief updating, planning, and acting, and gives practicable algorithms. It is the most frequently cited representative work for the "partially observable" situation, and Chapter 5's development of this face takes it as its base text; from it the reader can come to a systematic grasp of the whole practice of inference plus probing.

Budget-Constrained [1][4], including bounded rationality and the anytime algorithm

19. F. H. Knight (1921). *Risk, Uncertainty and Profit.* Houghton Mifflin. [1][4] Knight here distinguishes "risk,"

which can be characterized by probability, from "uncertainty," which cannot be quantified at all, the latter being a situation with no reliable probability to rely on. This distinction is one of the intellectual starting points for the book's discussion of the unverifiable, and it reminds the reader that the difficulty of some situations lies not in computing too imprecisely but in the absence of even the probability needed to place a bet.

20. H. A. Simon (1955). "A Behavioral Model of Rational Choice." *The Quarterly Journal of Economics*, 69(1), 99-118. [1][4] Simon here proposes bounded rationality: a real actor is limited in computation, time, and information, and so, rather than seeking the global optimum, "stops when satisfied." This is the conceptual source of the "budget-constrained" face, and this chapter borrows it to make the point that many verifications are feasible in theory but must be discounted once they land on a bounded actor, thereby leading into the later line of thought about budget allocation.
21. M. Boddy & T. Dean (1989). "Solving Time-Dependent Planning Problems." *Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI)*. [2][4] This paper continues the authors' work on anytime algorithms, studying how to arrange planning when computation time itself is limited, so that the system can be interrupted at any moment and hand over the current best solution. It shares a source with the next entry, and this chapter cites this body of work to show that the core of responding to the "budget-constrained" face is to spend the limited time where the marginal return is highest.
22. T. Dean & M. Boddy (1988). "An Analysis of Time-Dependent Planning." *Proceedings of the 7th National Conference on Artificial Intelligence (AAAI)*, 49-54. [2][4] This paper formally proposes the concept of the anytime

algorithm: the algorithm can be interrupted at any moment and yield the current best solution, with quality improving steadily as computation time increases. It is the representative formalization of the "budget-constrained" situation, and this chapter cites it to show that the distinctive feature of this face is that it recedes as resources grow, so that the key to responding lies in allocation rather than in raw computation.

23. S. J. Russell & D. Subramanian (1995). "Provably Bounded-Optimal Agents." *Journal of Artificial Intelligence Research*, 2, 575-609. [2][4] This paper formalizes bounded rationality as "bounded optimality": rather than requiring the agent to output the optimal decision, it requires the agent to do the best it can under a given constraint on computational resources. It gives Simon's intuition a precise definition, and this chapter cites it to show that the "budget-constrained" situation too can be theorized seriously, rather than being merely a helpless compromise.

Adversarial [2][1][4], including decision theory and adversarial machine learning

24. J. von Neumann & O. Morgenstern (1944). *Theory of Games and Economic Behavior*. Princeton University Press. [2][1] This book founds game theory, treating the interaction of many parties under conflicting interests as an object of rigorous analysis, and systematizes the minimax theorem for zero-sum games. It is the theoretical source of the "adversarial" face, and this chapter uses it to support the core claim of the face: in facing an opponent who optimizes against you, defend against the worst case rather than seek the optimum on some fixed input.

25. J. F. Nash (1950). "Equilibrium Points in N-Person Games." *Proceedings of the National Academy of Sciences*, 36(1), 48-

49. [2] Nash proves in this short paper that any finite many-person game has an equilibrium point, that is, a stable configuration in which no party can profit by unilaterally changing its strategy. The Nash equilibrium extends game analysis from the zero-sum to the general case, and this chapter cites it as a core concept of the "adversarial" face, helping the reader understand how strategic interaction converges to a predictable, stable structure.

26. A. Wald (1945). "Statistical Decision Functions Which Minimize the Maximum Risk." *Annals of Mathematics*, 46(2), 265-280. [2] Wald here founds statistical decision theory, proposing that decisions be chosen by the minimax criterion, that is, so as to minimize risk in the worst case. It carries "defend against the worst case" from games into statistical inference, and this chapter cites it to show that the cure for the adversarial face is a strategic posture: when the opponent will adjust to track your judgment, seeking robustness matters more than seeking the optimum at some single point.

27. L. J. Savage (1954). *The Foundations of Statistics*. Wiley. [2][1] Savage here builds the axiomatic foundations of subjective expected utility theory, arguing that a rational actor's preferences can be represented as the expected utility taken against a subjective probability. It is the standard framework for decision under uncertainty, and this chapter cites it to represent the orthodox stance of "keeping accounts of uncertainty with probability and utility," while also paving the way for the next entry to reveal the boundary of that stance.

28. D. Ellsberg (1961). "Risk, Ambiguity, and the Savage Axioms." *The Quarterly Journal of Economics*, 75(4), 643-669. [1][4] Ellsberg, with a simple betting experiment, reveals that people generally avoid the "ambiguous" option whose probability is itself unclear, a behavior that violates Sav-

age's axioms. It corroborates Knight's distinction between risk and uncertainty at the empirical level, and this chapter cites it to show that the probability framework is not omnipotent: some unverifiable situations cannot even yield a probability.

29. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow & R. Fergus (2014). "Intriguing Properties of Neural Networks." *International Conference on Learning Representations (ICLR)*. arXiv:1312.6199. [2] This paper is the first to reveal adversarial examples systematically: applying a perturbation almost imperceptible to the human eye to an image can make a neural network of extremely high recognition rate err. It brings the "adversarial" face into modern machine learning, and this chapter cites it to show that, so long as someone goes looking specifically for that perturbation, even the most accurate model can be fooled, which is exactly where the adversarial differs from a mere observability gap.
30. I. J. Goodfellow, J. Shlens & C. Szegedy (2015). "Explaining and Harnessing Adversarial Examples." *International Conference on Learning Representations (ICLR)*. arXiv:1412.6572. [2] This paper attributes adversarial examples to the model's linearity in high-dimensional space, proposes the FGSM method for rapidly generating perturbations, and defends against them with adversarial training. It both explains why adversarial examples are pervasive and gives the earliest means of response, and this chapter cites it to show that the attack and defense of the adversarial face are a pair locked in mutual rise and fall, demanding a continual game.
31. A. Madry, A. Makelov, L. Schmidt, D. Tsipras & A. Vladu (2018). "Towards Deep Learning Models Resistant to Adversarial Attacks." *International Conference on Learning Representations (ICLR)*. arXiv:1706.06083. [2][4] Madry and

colleagues write adversarial robustness as a minimax optimization problem: the inner layer finds the worst perturbation, the outer layer trains to resist it, with projected gradient descent as the standard attack. It unifies attack and defense in the language of robust optimization, landing this chapter's claims about the adversarial face and worst-case decision squarely within machine learning, and is a deeply influential paper in this direction.

32. B. Biggio & F. Roli (2018). "Wild Patterns: Ten Years after the Rise of Adversarial Machine Learning." *Pattern Recognition*, 84, 317-331. [2] This survey reviews a decade of adversarial machine learning, noting that the field got under way before deep learning became fashionable, and lays out the overall thread of attack models, threat modeling, and defense. For the reader who wishes to grasp the "adversarial" face in full, it is an authoritative overview, and this chapter cites it as the best foothold for reading the face through.

Chapter 3: Falsifiable, Not Verifiable

> **Thesis:** The most disciplined way humans have to seek knowledge, empirical science, rests on a public admission: a theory can never be verified, only left unfalsified.

The previous chapter asked whether humans have a mature, disciplined way to live alongside the unverifiable for the long haul. There is one: empirical science. And the most surprising thing about it is that its first principle is not the claim that it can establish the truth, but precisely the public admission that it can never do so.

A Black Swan

"All swans are white." You have seen a thousand white swans, you have seen a million, and this universal statement is still not verified, because the next one may be black. But you need only see a single black swan, and it is overturned completely. This is no made-up example: in Europe, "all swans are white" was long taken as unquestioned common sense, until in 1697 a Dutch expedition saw a black swan for the first time in Western Australia, and that "certainty" was voided overnight.

This asymmetry is the pivot of the whole chapter. To verify a universal proposition you must check every case it asserts, and those cases are typically infinite, open, and lodged in the future, so the thing simply cannot be done. But to falsify it, a single counterexample suffices. Written out in logic: $\forall x\, P(x)$ cannot be established by any finite set of observations, but a single $\exists x\, \neg P(x)$ is enough to shatter it. The entire discipline of science is built on recognizing and exploiting this asymmetry.

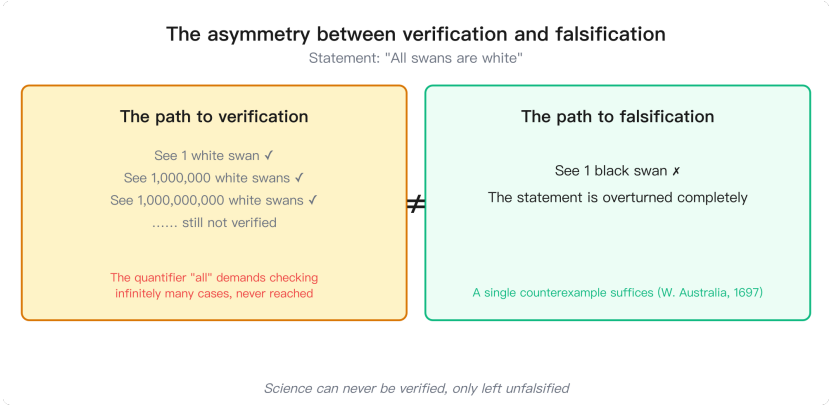

Figure 3: The asymmetry between verification and falsification

Hume's Impassable Threshold

The root of this was already dug out by Hume in 1739[4]. On what grounds do we believe that a regularity that has held in the past will go on holding in the future? There is no logical warrant for it. From "the sun has risen every day in the past" you cannot infer "the sun must rise tomorrow," because the inference itself presupposes that "past patterns will continue into the future," which is exactly the thing to be proved. Induction carries no logical guarantee. Hume's conclusion is calm and complete: what we rely on is not proof, but habit.

This is the philosophical footing of the "future" breach from Chap-

ter 1. Any knowledge of the world's general regularities is built on a finite past, and therefore cannot be verified in advance. If science is to count as knowledge, it cannot take "verification" as its goal, for that goal is simply out of reach.

Popper: Trading the Unreachable for the Reachable

Popper, in 1934[1] (in the original German), offered a way out: since verification is unattainable, stop demanding it, and use falsification instead. Whether a theory is scientific does not turn on how much evidence can support it (support can always be found), but on whether it has stuck its neck out and made risky predictions that could be overturned. Astrology, and any doctrine that explains away everything, is unfalsifiable and therefore unscientific; general relativity predicted that starlight would be bent by the sun's gravity through a specific angle, and the 1919 eclipse observation could perfectly well have measured a different value and so refuted the prediction. It is precisely because the theory dared to risk being overturned that it is good science.

So science became a machine optimized specifically for living alongside the unverifiable. It never claims to have proved anything; it says only that this theory has not yet been falsified, so we shall use it for now. This is a posture, a posture of trading the unmeasurable "true" for the measurable "not yet overturned." Look familiar? This is exactly what the mathematician's proxy substitution in Chapter 7 looks like at the scale of epistemology.

A Candid Qualification

It must be said plainly, right here: Popperian falsificationism is far from settled in the philosophy of science, and this book treats it as a clear point of entry, not as a final word.

Its most forceful objection comes from the holism of Duhem and Quine (also called the Quine-Duhem thesis). Duhem in 1906[10] and Quine in 1951[9] pointed out that you can never test a hypothesis in isolation. Any prediction depends on a large bundle of auxiliary assumptions (the instrument is not broken, the background conditions hold, the approximation is reasonable), and once an experiment fails, you can always turn the blame toward some auxiliary assumption and keep the core hypothesis intact. So the picture in which "a single counterexample cleanly overturns a theory" is not as clean as it looks. Kuhn in 1962[5] went further: scientists in a period of normal science are in no hurry to falsify at all; anomalies are set aside until a paradigm crisis brings about a revolutionary replacement. Lakatos in 1970[6] used the progress and degeneration of a "research programme" to replace black-and-white falsification; Feyerabend[7] simply opposed any unified method at all. Another route is Bayesian confirmation theory[24], which wants no two-valued verdict but instead treats evidence as an adjustment to the probability of a belief,

$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)},$$

which in turn foreshadows the later move of calibration. Mayo's "severe testing"[14] is a refined heir of falsificationism, while Stanford[19] reminds us that a great many "unconceived alternatives" still lie beyond our view.

Laying these disputes out does not dismantle Popper; it enacts what this book itself ought to do: state a powerful framework while marking its boundaries without flinching. This posture is exactly the move the whole book sets out to rehearse.

Science Discovered Those Moves Long Ago

Here is this chapter's real gift to the book. If you look at the everyday machinery of science with the eye of the eight moves, you find that it worked several of them out long ago, only under different names.

Peer review is redundancy and consensus: it distrusts any single judge and uses several mutually independent reviewers, taking their agreement. Replication is also redundancy: a result is not taken seriously until someone else reproduces it independently elsewhere. Preregistration is an audit trail: before the data are seen, the hypothesis and the analysis plan are registered, so the target cannot be moved afterward and noise cannot be talked up into signal. Confidence intervals and error statistics are certificates and bounds: they do not claim a proposition is true, only that, at an explicit confidence level, a bounded guarantee holds. Double-blinding and randomization are defenses against the fifth face (adversarial), and here the adversary is often the researcher's own bias and subjective expectation. The significance threshold is a crude form of calibration.

In other words, humanity's most serious enterprise of inquiry is itself a living sample of this book's convergence claim. This is the book's first and quite weighty hint: although the unverifiability science faces (about general regularities, about the future) has its own particular source, the responses it is forced into rhyme with the responses in software, mathematics, and organizations.

When the Machine Fails: The Replication Crisis

The reverse view makes it clearer. When these moves are weakened, science's self-correction breaks down, and that is the replication crisis. Ioannidis, in his 2005 paper[28] "Why Most Published

Research Findings Are False," and the Open Science Collaboration's 2015 large-scale replication[29] of a hundred psychology studies (97 percent of which had originally reported significant results, of which only about thirty-six still held after redoing them, fewer than half) lay this scene bare: when preregistration is absent (the target can be moved afterward), sample sizes are too small, publication bias passes only the pretty results, and few people undertake the thankless work of replication, the machine spins in neutral.

The diagnosis and repair of this crisis are carried out in the very language of those moves: restoring preregistration (putting the audit trail back), encouraging and rewarding replication (putting the redundancy back), registered reports, and raising the severity of testing. Problem and remedy fall onto the same vocabulary. This point will return in Chapter 10 on borrowed judgment and Chapter 12 on audit trails and auditing.

Where This Chapter Leads

Science has proved one thing: a person can seek knowledge in a disciplined way in a world without verification, and wherever this is done well, it relies on those very moves. This is a proof of concept for the whole book.

But a trap lurks here too. Precisely because all five faces appear with the same expression, "I cannot check it," and precisely because the moves for handling them are so similar, a deeply tempting thought arises: why not simply declare that unverifiability is one problem with one unified solution? The part of this thought about the "problem" is wrong; the part about the "response" stumbles onto something right. The next chapter is devoted to this temptation.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

1. K. Popper (1959). *The Logic of Scientific Discovery.* Hutchinson. [2][3] Popper here sets out falsificationism systematically: a scientific theory cannot be empirically verified, only refuted, so falsifiability becomes the line between science and non-science. The original German edition, *Logik der Forschung*, was published by Springer in Vienna, with the copyright page marked 1935 though it actually appeared in late 1934 (hence often dated 1934); this English edition was substantially revised and expanded by the author himself. The section "Popper: Trading the Unreachable for the Reachable" is built directly on this, and the reader should attend above all to the epistemological posture of replacing "verification" with "not yet falsified."

2. K. Popper (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge.* Routledge and Kegan Paul. [3] This essay collection unfolds falsificationism into a whole view of the growth of knowledge: knowledge advances through bold conjecture and merciless refutation, and the growth of science is not the accumulation of confirmations but the ceaseless weeding out of errors. Compared with the logical skeleton of the earlier work, it shows more vividly how "trial and error" drives scientific progress, and is good reading for understanding this chapter's waypoint of how science progresses.

3. K. Popper (1972). *Objective Knowledge: An Evolutionary Approach.* Clarendon Press. [3] Popper here likens the growth of knowledge to an evolutionary process of trial and

error, and proposes a "third world," the domain of objective knowledge itself, existing independently of any individual subjective mind. It pushes falsificationism toward an ontological picture of how objective knowledge can accumulate without a subject, and offers further reading for those who wish to pursue how science progresses in depth.

4. D. Hume (1739). *A Treatise of Human Nature*. John Noon. [2] Hume here raises the problem of induction, that source-level difficulty: from a past regularity one cannot infer a future regularity, because the inference itself presupposes the "uniformity of nature," which is precisely what is to be proved; our belief in causation and regularity comes, in the end, from habit rather than proof. Books I and II were published by John Noon in 1739, and Book III, *Of Morals*, by Thomas Longman in 1740, with the first edition conventionally dated 1739. The section "Hume's Impassable Threshold" is founded on this, the philosophical footing for understanding why science cannot take "verification" as its goal.

5. T. Kuhn (1962). *The Structure of Scientific Revolutions*. University of Chicago Press. [1][3] Kuhn, drawing on a great many cases from the history of science, argues that science does not approach truth at a steady pace, but solves puzzles within a shared paradigm during periods of "normal science," and only after anomalies accumulate into crisis does a paradigm-shifting scientific revolution occur, with old and new paradigms being incommensurable. It is an important correction to Popper's picture, showing that scientists are often in no hurry to falsify an anomaly. This chapter's "A Candid Qualification" cites it to mark the boundary of falsificationism.

6. I. Lakatos (1970). "Falsification and the Methodology of

Scientific Research Programmes." In I. Lakatos and A. Musgrave (eds.), *Criticism and the Growth of Knowledge*, pp. 91-196. Cambridge University Press. [1][3] Lakatos uses the "research programme" to reconcile Popper and Kuhn: each programme has a protected hard core and a surrounding belt of adjustable auxiliary assumptions, and the standard of judgment is not a single counterexample but whether the programme as a whole is, over time, "progressing" (continuing to make and fulfill new predictions) or "degenerating" (busy only with patching after the fact). It replaces black-and-white falsification with a historical judgment about a programme's advance or retreat, a key reference when this chapter delimits the boundary of falsificationism.

7. P. Feyerabend (1975). *Against Method: Outline of an Anarchistic Theory of Knowledge*. New Left Books. [1][3][4] Feyerabend argues forcefully, through cases from the history of science such as Galileo, that there is no universally valid set of scientific methods, and that major advances often come precisely from breaking existing rules, hence his famous slogan "anything goes." It is the most radical opposition to a unified methodology, cited in this chapter to make plain that even a methodological claim as mild as "falsification" is rejected at the root by some.

8. C. G. Hempel (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. Free Press. [2] In this essay collection Hempel gives a culminating account of the covering-law model of scientific explanation, including both deductive-nomological and inductive-statistical explanation, and discusses the logic of confirmation and its paradoxes. It represents logical empiricism's systematic characterization of the "theoretically studied material," and supplies this chapter with a classic background on what counts as testable and explicable.

9. W. V. O. Quine (1951). "Two Dogmas of Empiricism." *The Philosophical Review*, 60(1), 20-43. [2] Quine attacks the two dogmas of logical empiricism, the sharp divide between analytic and synthetic and reductionism, and proposes epistemological holism, holding that our beliefs face experience together as one whole web, and that no single statement can be verified or refuted in isolation. Combined with Duhem's holism of testing (jointly called the Quine-Duhem thesis), it strikes directly at the picture in which "a single counterexample cleanly overturns a hypothesis," a core reference for this chapter's delimiting of the boundary of falsificationism.

10. P. Duhem (1906). *La théorie physique: son objet, sa structure*. Chevalier & Rivière. [2] Duhem here proposes the holism of testing: an experiment in physics never tests an isolated hypothesis but rather "the hypothesis together with a whole set of auxiliary assumptions and background theories," so a failed prediction cannot determine exactly where the error lies. This is the source of what was later, with Quine, called holism, used in this chapter to show that the bearing of a counterexample is not as definite as it appears. The original 1906 French edition is taken as authoritative here; the second edition was published by Marcel Rivière in 1914, and P. P. Wiener's English translation, *The Aim and Structure of Physical Theory*, was issued by Princeton University Press in 1954.

11. M. Polanyi (1958). *Personal Knowledge: Towards a Post-Critical Philosophy*. University of Chicago Press. [1][4] Polanyi proposes "tacit knowledge": we know far more than we can tell, and in scientific inquiry there is always a layer of personal judgment and craft that cannot be formalized and can only be acquired through practice and apprenticeship. It reminds us that no methodology, however strict, can eliminate the scientist's own, inarticulable judgment, echo-

ing this chapter's waypoints of historical scientific judgment and how to live in an unverifiable world.

12. B. C. van Fraassen (1980). *The Scientific Image*. Clarendon Press. [2][4] Van Fraassen proposes "constructive empiricism": the aim of science is not to claim a theory is true but only that it is "empirically adequate," that is, that it correctly saves the observable phenomena; to accept a theory is to believe it empirically adequate, not to believe its unobservable parts actually exist. It turns "unverifiable" into a mature scientific attitude, resonating with this chapter's posture of replacing "true" with "not yet overturned."

13. I. Hacking (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press. [2][3] Hacking shifts philosophical attention from "representing" to "intervening," arguing that the best defense of realism lies not in theory but in experiment: when we can stably manipulate electrons to probe other things, electrons are real ("if you can spray them, they are real"). It opens a new approach to scientific realism grounded in experimental practice, and reminds the reader that scientific progress likewise depends on hands-on intervention, not on theoretical testing alone.

14. D. G. Mayo (1996). *Error and the Growth of Experimental Knowledge*. University of Chicago Press. [3] Mayo proposes the philosophy of "error statistics": we have reason to accept a hypothesis only when it has passed a severe test, one that "would very probably have failed if the hypothesis were false." This makes Popper's spirit of falsification operational as a statistical testing procedure, and is a refined heir of falsificationism; this chapter's notion of "severe testing" comes from here (part of the Science and Its Conceptual Foundations series).

15. D. G. Mayo (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars.* Cambridge University Press. [3][4] Mayo here reconstructs statistical inference around "severity" as a unifying principle, attempting to get past the long-running "statistics wars" between frequentists and Bayesians, and on this basis responds to criticisms of significance testing in the replication crisis. It develops the programme of item 14 into a methodology facing contemporary statistical practice, and is especially apt for understanding how to use statistical evidence responsibly in an unverifiable world.

16. L. Laudan (1981). "A Confutation of Convergent Realism." *Philosophy of Science*, 48(1), 19-49. [1][2] Laudan lists a batch of theories from the history of science that were once successful (able to predict, able to explain) yet were finally abandoned, such as phlogiston and the ether, arguing that the inference "success implies truth" does not hold up, a forceful rebuttal to convergent realism often called the "pessimistic meta-induction." It shows that even empirically very successful theories need not be close to the truth, reinforcing this chapter's claim that science does not take "verifying the truth" as its goal.

17. L. Laudan (1977). *Progress and Its Problems: Towards a Theory of Scientific Growth.* University of California Press. [3] Laudan argues for measuring scientific progress by "problem-solving capacity" rather than approach to truth: whether a research tradition progresses turns on the net gain in the empirical and conceptual problems it solves. It offers a view of progress that bypasses the concept of truth, supplying this chapter's how-science-progresses with an alternative framework that does not depend on verification.

18. P. Kitcher (1993). *The Advancement of Science: Science*

without Legend, Objectivity without Illusions. Oxford University Press. [3] Kitcher, having discarded the "legend" of science as all-knowing and all-powerful, also refuses relativism, and instead rebuilds a moderate and defensible objectivity and view of progress from science's social and cognitive practice. It demonstrates how one can, while admitting that science is shaped by history and society, still hold onto the two concepts of progress and objectivity, in line with this chapter's stance of affirming science while honestly marking its boundaries.

19. P. K. Stanford (2006). *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives.* Oxford University Press. [1][2][3][4] Stanford raises the problem of "unconceived alternatives": the history of science shows again and again that past scientists always had theoretical options that only appeared later and were utterly unthinkable at the time, so we have no reason to believe that we have today exhausted all viable explanations. He distills this "new induction" from historical cases such as genetics, directly echoing this book's framework of "an unverifiable world," and reminding the reader that there is always an unconceived possibility beyond our view.

20. N. Goodman (1955). *Fact, Fiction, and Forecast.* Harvard University Press. [2] Goodman raises the "new riddle of induction": "green" and the artificial predicate "grue" (meaning observed as green before a certain time and blue thereafter) both fit all observations to date equally well, yet yield opposite predictions, which shows that induction cannot be settled by evidence alone but must also depend on which predicates are "projectible." It shows that the difficulty of induction is not only the Humean problem of justification but, more deeply, the indeterminacy of the regularity itself, deepening this chapter's understanding of why induction is

unreliable. The first-edition year is commonly given as 1955 (one HUP blurb gives 1954, a slight ambiguity; the widely cited 1955 is followed here).

21. C. G. Hempel and P. Oppenheim (1948). "Studies in the Logic of Explanation." *Philosophy of Science*, 15(2), 135-175. [2] Hempel and Oppenheim here lay the foundation of the deductive-nomological (D-N) model of explanation: that a phenomenon is scientifically explained means that it can be logically derived from general laws plus initial conditions. It is the starting point of twentieth-century theories of scientific explanation, defining what "being explicable" means logically, and supplying this chapter with an underlying framework for how science characterizes regularities.

22. R. Carnap (1936-1937). "Testability and Meaning." *Philosophy of Science*, 3(4), 419-471; 4(1), 1-40. [2] Carnap here loosens the strict principle of verifiability, using the broader notions of "testability" and "confirmability" to demarcate meaningful empirical statements, and handles the connection between theoretical terms and observation through devices such as disposition predicates. It records logical empiricism's key retreat from "verifiable" to "testable," resonating exactly with this chapter's main line of "verification is out of reach, so use the reachable instead." The text was published in two parts, volume 3 issue 4 (1936) and volume 4 issue 1 (1937).

23. W. C. Salmon (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press. [2] Salmon argues that the heart of scientific explanation is not logical derivation but the disclosure of causal mechanisms: to explain a phenomenon is to embed it in the world's web of causal processes and causal interactions. It is an important correction to the covering-law model, shifting the

standard of "being explicable" from derivability to traceable causal structure, supplying this chapter's understanding of how science characterizes the world with the dimension of causation.

24. C. Howson and P. Urbach (1989). *Scientific Reasoning: The Bayesian Approach*. Open Court. [2][3] Howson and Urbach systematically advocate a Bayesian view of scientific reasoning: rather than a true-or-false two-valued verdict, they treat evidence as a continuous adjustment, by Bayes's theorem, to the probability of a belief, and on this basis respond to many difficulties of induction and confirmation. It is the representative work on the Bayesian confirmation theory mentioned in this chapter's text, forming a contrast with falsification and severe testing, and foreshadowing the book's move of calibration.

25. E. Sober (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge University Press. [2] Sober uses the tools of likelihood theory and statistical inference to analyze in detail "what evidence supports," and discusses which hypotheses are genuinely testable, including an anatomy of why intelligent design is untestable. It brings the abstract question of testability down to concrete scientific inferential practice (especially with evolution as the example), demonstrating how to judge rigorously whether a claim can withstand the test of evidence.

26. P. Godfrey-Smith (2003). *Theory and Reality: An Introduction to the Philosophy of Science*. University of Chicago Press. [2][3] Godfrey-Smith's widely praised introduction to the philosophy of science lays out clearly the whole thread from logical empiricism, falsificationism, and Kuhn's paradigm to Bayesianism and the dispute over scientific realism. It serves well as an introductory anchor for the

many topics of this chapter, and the reader who wants to build a global map before reading the monographs can begin here.

27. N. Cartwright (1983). *How the Laws of Physics Lie.* Clarendon Press. [2][3] Cartwright argues that the fundamental laws of physics are universal and elegant precisely because they do not faithfully describe the real world but rather describe highly idealized models; the more fundamental a law, the greater its explanatory power, yet the more it "lies" in description. She turns instead to value the concrete laws and causal capacities closer to the phenomena, reminding the reader that the "truth" of scientific laws is far more complex than usually supposed, deepening this chapter's reflection on the relation between theory and world.

28. J. P. A. Ioannidis (2005). "Why Most Published Research Findings Are False." *PLoS Medicine*, 2(8), e124. [3][4] Ioannidis argues, through concise statistical modeling, that in fields with small effect sizes, flexible study designs, and rampant publication bias, the probability that a published "positive" finding is false is often higher than the probability that it is true, and false positives can systematically outnumber true ones. It is a founding document of the replication crisis, and the core evidence for this chapter's section "When the Machine Fails," showing how science's self-correction spins in neutral once it is weakened.

29. Open Science Collaboration (2015). "Estimating the Reproducibility of Psychological Science." *Science*, 349(6251), aac4716. [3][4] The Open Science Collaboration, coordinating over a hundred researchers, conducted systematic direct replications of a hundred published psychology studies, with the result that fewer than half successfully reproduced the original effect, and the reproduced effects were generally

weaker than originally reported. It turns the replication crisis from argument into large-scale evidence, the empirical core of this chapter's "replication crisis" section, and underscores why moves such as replication and preregistration are indispensable.

Chapter 4: The Temptation to Flatten

> **Thesis:** Because all five faces present as "I cannot check it," there is a strong urge to treat unverifiability as one problem and fit it with one solution. About the problem this is wrong, yet it points toward the right thing about the response.

Chapter 2 spent a whole chapter prying unverifiability apart into five faces. Chapter 3 then observed that the moves for coping with those faces resemble one another. Put the two together, and a thought almost inevitably arises, and a deeply tempting one: since they all carry the same feature, "I cannot check it," and are all handled by much the same means, why not simply declare that unverifiability is one problem and fit it with one unified solution?

This chapter does two things. First it exposes where that thought goes wrong; then it makes clear what, by accident, it gets right, and from that draws for the whole book a burden of proof.

Where Flattening Goes Wrong

To flatten is to crush five structurally different faces into a single face. Its cost was, in fact, already rehearsed in Chapter 2.

To treat the undecidable as a budget problem, supposing that "throwing in more computing power will do it." Wrong. The halting problem is not slow to compute; there is simply no such algorithm, and no faster machine can conjure a program that does not exist.

To treat an adversarial gap as a mere observability gap, supposing that "I just have not seen it clearly yet." Wrong, and more dangerously so. The system across from you adjusts itself in light of how you look: the more clearly you see, the more cunningly it hides. This is a game of chess, not a single measurement. Spam filtering is the living textbook: you train a classifier on the features of today's spam, and the senders immediately rewrite their wording, switch domains, and insert garbled characters to slip past. Whoever mistakes it for a static recognition task and solves it with a one-off model will forever be half a step behind.

To treat the intractable as the undecidable, and so abandon too early a problem that could in fact be approximated or handled in the average case; or the reverse, to treat the undecidable as merely intractable, and pound away with computing power against a wall of logic without end. Every such misidentification makes you reach for the wrong tool and pay a real price. Diagnose the lesion wrongly, and even the most apt medicine is poison.

So, about the "problem," flattening is wrong. The five faces must be handled separately. This is the conclusion Chapter 2 bought with a whole chapter, and it cannot be lightly handed back here.

The Side Flattening Accidentally Gets Right

And yet there is a grain of truth inside that thought. About the "problem" it is wrong; about the "response" it accidentally gets it right.

This book's thesis is precisely the picking of that grain of truth

out of the surrounding error, and stated with great care:

> The sources of unverifiability differ wildly, but the responses a bounded actor is forced into converge again and again on the same small set.

Note the restraint of this formulation. It does not say "these problems are the same problem." That would be flattening, and wrong; any expert in any one field would throw the book down. It says something else, stronger and more defensible: the problems all differ, but the responses rhyme. What the book sets out to deliver is that "table of the same move under many vocabularies," together with an explanation of why the convergence falls on just these few moves.

A Burden of Proof It Must Bind Itself With

Said this far, the greatest risk surfaces too. The human mind is born to love analogy. Lakoff and Johnson[14] let us see that even everyday language is built out of metaphor, while Gentner[6], Holyoak[9], Bartha[7], and others have studied when an analogy is true and when it is merely good-looking. But precisely because analogy comes so readily to hand, it is also the easiest to be fooled by. A beautiful cross-domain analogy often proves nothing at all. The handiest cautionary case is the lineage of Hofstadter's *Gödel, Escher, Bach*[1]: cross-domain resonances written so dazzlingly that they take the breath away, yet often criticized as "in the end just analogy," unable to survive a hard question. A more solid lesson comes from the so-called power law craze. Many systems were declared to obey the same power law and share the same deep mechanism, which sounded marvelously unified; yet once tested with statistics as strict as those of Clauset, Shalizi, and Newman[25] in 2009, a great many "power laws" simply failed to hold. They re-examined more than twenty real data sets widely claimed to be power laws, and only a handful passed the test

cleanly; most were in fact better fit by other distributions such as the log-normal. Stumpf and Porter[26], in their 2012 piece "Critical Truths About Power Laws," put it bluntly: looking like one is not the same as being one.

So this book must bind itself with an iron rule: any claimed convergence must be shown to be more than analogy.

What counts as "more than analogy"? The philosophy of science offers a ready measuring stick. For a cross-domain mapping to count as substantive, it must be structure-preserving, not merely surface-similar but corresponding in mechanism, corresponding in mode of failure, corresponding in trade-off. Gentner's[6] structure-mapping and Bartha's[7] evaluation of the analogical argument give exactly this set of standards. There is a still harder criterion, robustness: a conclusion is more credible if it can be derived again and again from several mutually independent routes. The robustness analysis developed by Levins[17], Wimsatt[18], and Weisberg[19] is the very source of this idea. Anderson's[5] line "More Is Different," Fodor's[27] argument about the "special sciences," and Cartwright's[28] *The Dappled World* all show that real cross-level patterns do exist, but that they are earned, not declared. Box's[33] famous line hangs overhead: all models are wrong, but some are useful. What this book strives for is "useful, and useful in a way that survives testing," not "so elegant that one forgets to test it."

In operational terms, this iron rule means that every time Part III extracts a move and every time it sets two domains side by side, it must interrogate once more: is this transfer substantive (same mechanism, same mode of failure, same trade-off), or merely a pretty figure of speech? Survive that interrogation, and the table stands; fail it, and the table is only a well-made piece of prose. Chapter 13 will attempt to hang this convergence on a common underlying structure (the decomposition of risk and information), but that is a promissory note still to be honored, something to be

tested, not something that may be assumed in advance. Chapter 14 will then settle the account squarely: is this a law, or a very strong empirical pattern?

Where This Chapter Leads, and the Order of Part II

Since we cannot rely on declaring the convergence in advance and then picking a few examples to fit it, the only sound way to test it is to walk into real domains, see what capable actors actually did, and let the moves grow out of the cases themselves rather than fixing the moves first and then forcing the cases to match.

This also explains why the book places the sites (Part II) before the toolbox of moves (Part III). To abstract first and exemplify later would seem arbitrary, and would waste the persuasive force the cases ought to carry. So Part II does not rush to name; it lets the moves appear embedded in their own domains, intertwined with one another, looking somewhat messier on the surface. Only in Part III is each move lifted out on its own, cleaned, and named. Induction first, naming after.

Four sites are ready: a designer feeling out what a user has in mind, an agent set loose to act on its own, a mathematician beating against the Riemann hypothesis, and a vast organization that cannot see itself. The unverifiability they face comes from different ones among the five faces. Let us go and see what each of them reaches for, when the oracle never comes.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses;

4. how to live in an unverifiable world. This section was checked source by source.

1. D. Hofstadter (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books. [2] Hofstadter, drawing on Gödel's incompleteness, Escher's visual paradoxes, and Bach's fugues, weaves a grand analogy about how self-reference, recursion, and consciousness emerge from formal systems. It is a benchmark of cross-domain analogical writing, and also this chapter's cautionary case: the resonances are written so dazzlingly that they take the breath away, yet have repeatedly been criticized as "in the end just analogy," a reminder that a beautiful cross-domain resonance does not itself constitute an argument.

2. H. A. Simon (1969). *The Sciences of the Artificial*. MIT Press. [2][3] Simon proposes a science of the "artificial," arguing that design is a subject that can be made systematic, and characterizes the decisions of real actors through bounded rationality and satisficing rather than optimization. Its importance to this chapter lies in treating "how a bounded actor copes under constraints" as a proper object of science, which is the very source of the "responses forced out of an actor" in the book's thesis.

3. W. C. Wimsatt (2007). *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Harvard University Press. [2][4] Wimsatt argues that philosophy should be rewritten for "limited beings": actors with finite cognitive resources can only approximate reality through heuristics, approximations, and robustness, and error is inevitable yet manageable. The title is almost a footnote to this book's secondary thread, and the reader may focus on how it treats robustness as the central tool by which a bounded actor tells real patterns from false ones.

4. H. A. Simon (1962). "The Architecture of Complexity." *Proceedings of the American Philosophical Society*, 106(6), 467-482. [2][3] Simon argues that complex systems are mostly "near-decomposable" hierarchical structures, with tight connections inside subsystems and loose connections between them, an architecture that eases both evolution and understanding. It provides a classic structural argument that "real patterns exist across levels," worth reading alongside Anderson and Fodor.

5. P. W. Anderson (1972). "More Is Different." *Science*, 177(4047), 393-396. [2] Anderson opposes the arrogance of reductionism, pointing out that each level brings forth new regularities that the laws of the level below cannot simply derive. This chapter's phrase "More Is Different" comes from here; it shows that real cross-level patterns do exist, but must be earned by their own sciences rather than declared from fundamental physics.

6. D. Gentner (1983). "Structure-Mapping: A Theoretical Framework for Analogy." *Cognitive Science*, 7(2), 155-170. [2] Gentner proposes the structure-mapping theory: a good analogy transfers relational structure rather than surface properties, and a systematic web of relations is more valuable than isolated points of similarity. This is the theoretical source of this chapter's criterion that "a substantive analogy must be structure-preserving," and the reader can understand from it what "corresponding in mechanism rather than surface similarity" means.

7. P. Bartha (2010). *By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments*. Oxford University Press. [2] Bartha builds a normative framework for evaluating analogical arguments, asking when an analogy can truly bear the weight of inference; the key is whether a relevant

causal or structural connection holds between the source domain and the target domain. It supplies an operable criterion for this chapter's iron rule of "more than analogy."

8. M. B. Hesse (1966). *Models and Analogies in Science*. University of Notre Dame Press. (Expanded edition; first published 1963.) [2][3] Hesse analyzes the cognitive functions of models and analogies in science, distinguishing positive, negative, and neutral analogies, and pointing out that the "neutral part" of an analogy is exactly the growth point for scientific prediction and discovery. It is an early founding work of the methodology of analogy, paving the way for the later structure-mapping and analogy evaluation.

9. K. J. Holyoak and P. Thagard (1995). *Mental Leaps: Analogy in Creative Thought*. MIT Press. [2] Holyoak and Thagard propose a multiconstraint theory of analogy, holding that the human mind completes an analogical mapping under the mutual balancing of three kinds of constraint: structural, semantic, and pragmatic. The book examines analogy within the real processes of cognition and creation, helping the reader understand why analogy is both powerful and error-prone.

10. D. Gentner, K. J. Holyoak, and B. N. Kokinov (Eds.) (2001). *The Analogical Mind: Perspectives from Cognitive Science*. MIT Press. [2] This collection gathers research on analogy from across the cognitive sciences, from computational models to developmental psychology to neural mechanisms, systematically presenting the research landscape of "when an analogy is true and when it is merely good-looking." It is the chapter's overview entry into the lineage of research on analogy.

11. D. Hofstadter and E. Sander (2013). *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. Basic Books. [2]

Hofstadter and Sander argue that analogy is the central engine of thought, that even the most basic categorization and concept formation are ongoing analogy. The book pushes the standing of analogy to its limit, which forms a tension with this chapter's wariness: analogy is everywhere, and precisely for that reason a set of criteria is all the more needed to tell which transfers are substantive.

12. S. Vosniadou and A. Ortony (Eds.) (1989). *Similarity and Analogical Reasoning.* Cambridge University Press. [2] This collection concentrates on the relation between similarity and analogical reasoning, asking what "similarity" really means and how it drives reasoning and learning. It provides conceptual preparation for the question behind this chapter: how to tell "looking like" apart from "actually being."

13. K. Dunbar (1995). "How Scientists Really Reason: Scientific Reasoning in Real-World Laboratories." In R. J. Sternberg and J. E. Davidson (Eds.), *The Nature of Insight*, 365-395. MIT Press. [1][2][3] Dunbar observed molecular biology laboratories in the field and found that scientists make heavy use of analogy in real work, and that near, within-domain analogies are often more fruitful than distant ones. With field evidence it supports this chapter's methodological choice to "let the moves grow out of the cases themselves," showing that real reasoning differs from the textbook account.

14. G. Lakoff and M. Johnson (1980). *Metaphors We Live By.* University of Chicago Press. [2] Lakoff and Johnson argue that metaphor is not mere rhetoric but the very structure of the human conceptual system, and that even everyday expressions like "time is money" conceal systematic metaphors. This chapter cites it to show that analogy and metaphor are deeply rooted in human thought, and that precisely for this

reason their reliability calls for more caution.

15. A. Tversky and D. Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases." *Science*, 185(4157), 1124-1131. [2][4] Tversky and Kahneman reveal that people make judgments under uncertainty by relying on heuristics such as representativeness, availability, and anchoring, shortcuts that are efficient yet lead to systematic biases. It reminds the reader that a bounded actor's responses are often not optimal solutions but makeshift moves, and also explains why a beautiful analogy is so apt to fool us.

16. R. W. Batterman (2001). *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford University Press. (Hardcover first published November 2001; some catalogues record it as 2002.) [2][3] Batterman studies the role of asymptotic reasoning in explanation, arguing that the explanation of many physical phenomena hides precisely in the "details" that emerge when taking a limit, and that universality has its source there. It provides a fine philosophical case for "how a common structure across systems is possible," echoing this chapter's inquiry into the mechanism of convergence.

17. R. Levins (1966). "The Strategy of Model Building in Population Biology." *American Scientist*, 54(4), 421-431. [2][3] Levins points out that model building cannot at once attain generality, precision, and realism, that the modeler must trade off, and proposes that when several models with different assumptions yield a consistent conclusion, that conclusion is more credible. This is the source of robustness analysis, from which comes this chapter's hard criterion that "what is derived again and again from multiple routes is more credible."

18. W. C. Wimsatt (1981). "Robustness, Reliability, and

Overdetermination." In M. B. Brewer and B. E. Collins (Eds.), *Scientific Inquiry and the Social Sciences*, 124-163. Jossey-Bass. [2][3] Wimsatt systematically develops the concept of robustness: what can be jointly detected by several mutually independent means, models, or perspectives is more likely to be real than an artifact. This paper is the core source for this chapter's robustness criterion, and from it the reader can understand why the convergence of independent routes suppresses error.

19. M. Weisberg (2006). "Robustness Analysis." *Philosophy of Science*, 73(5), 730-742. [2][3] Weisberg re-clarifies the logic of robustness analysis, distinguishing a robust theorem from the test of its empirical adequacy, and clarifying when it can and cannot lend support to a conclusion. It makes this chapter's robustness criterion more precise, reminding the reader that robust does not automatically equal true, and that empirical checking is still required.

20. S. H. Orzack and E. Sober (1993). "A Critical Assessment of Levins's The Strategy of Model Building in Population Biology (1966)." *The Quarterly Review of Biology*, 68(4), 533-546. [2][3] Orzack and Sober critically examine Levins's modeling strategy, questioning whether the agreement of several models alone can logically warrant inferring a conclusion to be true, unless those models have each already received independent support. It is an important counterweight to the robustness argument, helping this chapter hold its iron rule more tightly and avoid mistaking agreement for proof.

21. N. Goldenfeld and L. P. Kadanoff (1999). "Simple Lessons from Complexity." *Science*, 284(5411), 87-89. [2][3] Goldenfeld and Kadanoff remind us that studying complex systems calls for using the right model at the right scale, that uni-

versality, however alluring, should not obscure the concrete mechanism, and that the key is to "make the right simplification at the right level." It provides this chapter a sober voice from within physics on how to treat cross-system universal regularities with care.

22. L. P. Kadanoff (1966). "Scaling Laws for Ising Models near Tc." *Physics Physique Fizika*, 2(6), 263-272. [2] Kadanoff proposes the block-spin scaling picture, showing that near the critical point a system is self-similar across different scales, and laying the foundation for the later renormalization group and the theory of universality classes. It is a classic example of the genuine universality in which "different systems share the same critical behavior," standing in exact contrast to the later "power laws" that fail to hold.

23. P. Bak, C. Tang, and K. Wiesenfeld (1987). "Self-Organized Criticality: An Explanation of the 1/f Noise." *Physical Review Letters*, 59(4), 381-384. [2] Bak, Tang, and Wiesenfeld propose self-organized criticality, using the sandpile model to show that certain systems evolve spontaneously to a critical state, giving rise to power-law distributions and 1/f noise. It set off the later power law craze, serving both as a representative of the cross-system unifying narrative and as an object of this chapter's call for "strict statistical testing."

24. A.-L. Barabási and R. Albert (1999). "Emergence of Scaling in Random Networks." *Science*, 286(5439), 509-512. [2][3] Barabási and Albert propose the scale-free network model, explaining the power-law degree distribution of many real networks through the mechanism of growth plus preferential attachment. It is a founding work of network science, and also belongs to that batch of systems widely claimed to obey the same power law, fit to be re-examined under this chapter's critical lens.

25. A. Clauset, C. R. Shalizi, and M. E. J. Newman (2009). "Power-Law Distributions in Empirical Data." *SIAM Review*, 51(4), 661-703. [2] Clauset, Shalizi, and Newman propose a rigorous statistical method to test whether data truly obey a power law, including maximum-likelihood fitting and comparison against alternative distributions. After re-checking with this method, many previously claimed "power laws" do not hold. It is exactly the methodological model for this chapter's iron rule that "any claimed convergence must survive strict testing."

26. M. P. H. Stumpf and M. A. Porter (2012). "Critical Truths About Power Laws." *Science*, 335(6069), 665-666. [2] Stumpf and Porter sum up the lessons of power-law research, stating bluntly that looking like a power law is far from being one, still less from there being a common deep mechanism behind it, and that statistical fit and mechanistic explanation must be kept apart. This chapter's phrase "looking like one is not the same as being one" derives from here, and is the direct basis for tightening the burden of proof.

27. J. A. Fodor (1974). "Special Sciences (or: The Disunity of Science as a Working Hypothesis)." *Synthese*, 28(2), 97-115. [2] Fodor argues that the laws of "special sciences" such as psychology and economics are multiply realizable and cannot be reduced to physics, so that science is in essence disunified. This chapter cites it to support that "real cross-level patterns do exist and cannot be declared away by reduction," in the same camp as Anderson and Cartwright.

28. N. Cartwright (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge University Press. [2][3] Cartwright argues that the world is "dappled," that physical laws hold only within their own local domains and do not

constitute a single unified picture covering everything, and that universality is the exception rather than the norm. The book lends strong metaphysical support to this chapter's stance that "real patterns are earned, not declared."

29. M. Mitchell (2009). *Complexity: A Guided Tour*. Oxford University Press. [2][3] Mitchell writes a clear and reliable guide to the science of complex systems, covering information, computation, evolution, networks, and emergence, while keeping a cautious distance from the field's common overstatements. It is a safe entry point for the reader into the topics of complexity and power laws, and in attitude it accords with this chapter's restraint.

30. D. Sornette (2006). *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder* (2nd ed.). Springer. (First edition 2000.) [2] Sornette systematically surveys the mathematics behind critical phenomena, power laws, fractals, and self-organization in the natural sciences, giving technical tools for handling such heavy-tailed and scaling behavior. It represents the ambition to find universal scaling laws across disciplines, and can be read alongside the literature critical of power laws to see the distance between claim and test.

31. G. West (2017). *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies*. Penguin Press. [2] West proposes that everything from organisms to cities to companies follows certain scaling laws, such as the sublinear scaling of metabolic rate with body size, attempting to find unified quantitative laws for life and society. The book is a contemporary representative of the grand cross-domain universality narrative, fit for the reader to weigh by this chapter's criteria: which are substantive convergences, and which only a

moving vision of unity.

32. P. Galison (1997). *Image and Logic: A Material Culture of Microphysics*. University of Chicago Press. [1][3] Galison studies the experimental culture of twentieth-century microphysics, proposing that different subdisciplines collaborate in a "trading zone" through a working pidgin, able to work together even when their theoretical frameworks do not agree. It demonstrates how different fields genuinely connect with one another without being forcibly unified, echoing this book's emphasis on "rhyming rather than being the same."

33. G. E. P. Box (1976). "Science and Statistics." *Journal of the American Statistical Association*, 71(356), 791-799. [3][4] Box here sets out science as an iterative process of repeatedly matching model against reality and gradually closing in, and leaves the famous line "all models are wrong, but some are useful." This chapter hangs that line overhead to define what the whole book strives for: useful, and useful in a way that survives testing, not so elegant that one forgets to test it.

34. T. S. Kuhn (1962). *The Structure of Scientific Revolutions*. University of Chicago Press. [1][3] Kuhn proposes the famous framework of paradigm and scientific revolution, distinguishing the puzzle-solving of normal science from the incommensurability of paradigm change, and reshaping the understanding of "how science progresses." It is a foundational work coloring this book's thinking about scientific progress and judgment, reminding the reader that comparison across paradigms is never a simple item-by-item correspondence.

Part II: Incarnations

Chapter 5: The Human at the Console

> **Thesis:** When what you must satisfy is a person's true preference or intent, you are facing permanent partial observability: the latent goal cannot be read out directly, and the capable response is to put a judging actor into the loop, and to question it sparingly and intelligently.

What You Want Is Not What You Said

A scene told to death, yet always true: the user describes what he wants, the engineer builds it to the letter, and on the day of delivery the user says, no, this is not what I wanted.

No one lied. The user spoke the truth, and the engineer did as told. The trouble lies deeper: the thing the user truly wanted was never, from the very start, fully sayable, and could not be. This chapter looks at what capable people do when the goal they must satisfy is locked inside another person's head. The unverifiability here belongs to the "partial observability" among the five faces of Chapter 2: the relevant state is hidden from you, and not hidden temporarily but permanently. You cannot read the goal out of a person's head, and so you cannot verify whether you have actually

satisfied it.

The Latent Preference

Let us state this precisely. The user's true preference is a latent variable. It drives his reactions yet never shows itself directly; you can only infer it obliquely from his behavior.

What makes it worse is that this latent goal often cannot be read out even by the user himself. The psychologist Slovic[11] has an unwelcome but solid claim: preferences, much of the time, are not expressed but constructed in the very moment of being asked. When you ask a person what he wants, the answer he gives you is usually shaped together by your phrasing, by the options at hand, and by whatever reference point he happened to think of, not drawn from some preexisting, well-defined store of preference. This means that the seemingly safe order of "first pin down the requirement, then build it" rests on an assumption that often fails: that the requirement, as a definite object, exists prior to the asking.

So what you face is not a situation of "information temporarily missing, fillable by topping it up." Even if the user cooperates throughout and tells you all he knows, the goal still cannot be measured precisely. This is the purest human form of partial observability.

Why Asking Once Is Not Enough

If a preference were a fixed target, then asking once, and asking clearly, would in principle suffice. It is not.

Economics long ago separated two things: stated preference (what a person says he wants) and revealed preference (what a person's actual choices expose him to want), and the two frequently fail

to agree. A requirements document is a lossy compression: it squeezes a living intent, one that shifts with circumstance, into a static list of items, and what gets discarded is precisely the things not thought of at the time but instantly pointable-to once the finished product appears. Intent itself drifts, too: after seeing a concrete implementation, a person's preference is recalibrated by that implementation, and what he wants now is no longer what he wanted when the project began.

So "asking once" fails not because you asked badly, but because the nature of this object means a single inquiry cannot lock it down. The only thing that can cope with it is one structure: act, observe, and correct, again and again.

The First Move: Put the Judge in the Loop

The first response is to admit that you cannot read the goal out, and so to bring in, at every decision point, the one actor who does know the goal, and let it correct your course. Act, observe the reaction, update, act again. Put the human in the loop.

This loop has been reinvented separately in many fields. In human factors, Sheridan's[12] "human supervisory control" positions the human as a judge who supervises and intervenes above the automation, not a role replaced once and for all by a specification. In usability engineering, the experiential wisdom of Gould and Lewis in 1985[20] compresses it into three principles so plain they almost sound like platitudes, yet are violated by countless projects: focus on users early and continuously, measure empirically, and design iteratively. Nielsen[19] later engineered this into a whole set of usability methods, and offered a surprising empirical figure: just five users in a test will surface about eighty-five percent of the usability problems, so rather than bring in twenty people at once, it is better to run four rounds of five, testing and fixing as you go. A recommender system learns the tastes a user

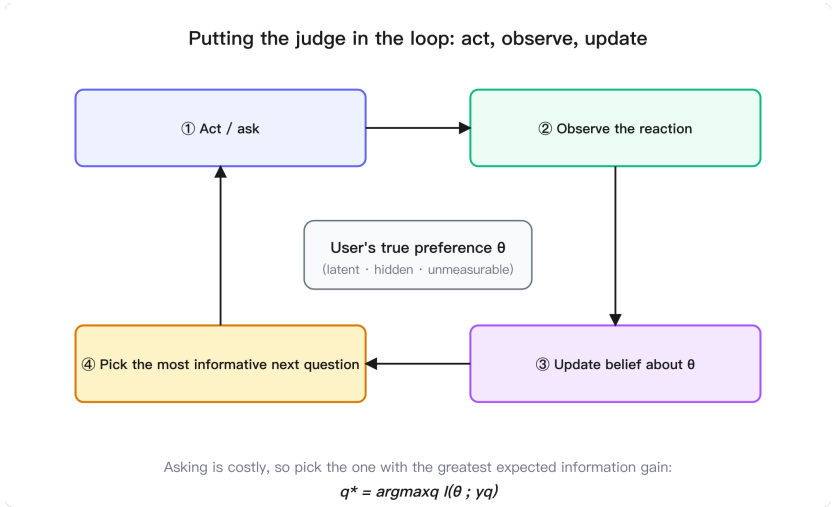

Figure 4: Putting the judge in the loop: act, observe, update

never voiced from his clicks, dwell times, and skips; at bottom this is the same loop. Horvitz's 1999[22] mixed-initiative user interface and the interactive machine learning proposed by Fails and Olsen in 2003[23] are both describing the same thing: human and system taking turns, calibrating one another.

Here a narrative collapse must be guarded against: interactive elicitation is not one specific technique, it is a family of methods. Experimental design, active learning, sequential decision, even exploration in reinforcement learning, are all instances of this same "act-observe-update" loop under different assumptions. To call it "just A/B testing" or "just some algorithm" would shrink a general posture down into a single tool.

The Second Move: Spend Each Question Where It Cuts Deepest

For the loop to turn, you must keep putting questions to the person, and asking has a cost. The user's patience, attention, and

time are all scarce; ask too much and too clumsily, and he will tire, give perfunctory answers, or walk away. Hence the second move: since checking has a cost, spend the limited supply of questions where the information is greatest.

This move has a clean theory. Lindley in 1956[1] gave a measure of the information an experiment provides, and Howard in 1966[3] proposed information value theory, turning "is it worth paying a cost to obtain this information" into a computable decision. Bayesian experimental design (the review by Chaloner and Verdinelli[5] is a good map) systematizes it: among all the questions you could ask, pick the one expected to compress your uncertainty the most. Formally, if θ is the latent preference you wish to infer and y_q is the answer to question q, you want the q that maximizes expected information gain:

$$q^\star = \arg\max_q \; \mathbb{E}_{y_q}\big[\,\mathrm{H}(\theta) - \mathrm{H}(\theta \mid y_q)\,\big] = \arg\max_q \; I(\theta; y_q),$$

that is, the q that maximizes the mutual information between the answer and the goal. In machine learning this idea is called active learning: the statistical active learning of Cohn and colleagues in 1996[6], the uncertainty sampling of Lewis and Gale in 1994, and the query by committee of Seung and colleagues in 1992[7] all ask the same question: which sample is the most worthwhile place to spend the next label. When a user finds it hard to assign a score yet easy to pick the better of two options, pairwise comparison (the Bradley-Terry model[2], $P(a \succ b) = \sigma(s_a - s_b)$) becomes one of the most information-efficient ways to ask.

The same collapse must be guarded against. The title of the 2016 review by Shahriari and colleagues is telling: "Taking the Human Out of the Loop," about using Gaussian processes for Bayesian optimization to automatically pick the next point to try. It is extremely useful, but it is only one implementation within this fam-

ily of methods, not the whole of "optimal screening." To equate this move with Gaussian processes is to equate transportation with the automobile.

The Contemporary Incarnation, and Its Backlash

Put these two moves together and you have today's mainstream method for aligning large models. Reinforcement learning from human feedback (RLHF; founded by Christiano and colleagues in 2017[27], applied to summarization by Stiennon and colleagues in 2020[28], and the InstructGPT of Ouyang and colleagues in 2022[29]) does exactly this: it uses people's pairwise comparisons to learn a reward model, then uses that model as a proxy for human preference to optimize the system. It stitches together "act-observe-update" and "spend each question where it cuts deepest." Its effectiveness is so striking that it is often punctured by a single comparison: an InstructGPT of only 1.3 billion parameters, fine-tuned on human feedback, had its outputs preferred by people over those of the original GPT-3, which was more than 100 times larger, at a full 175 billion parameters. Aligning with human preference sometimes matters more than simply piling the model bigger.

And its mode of failure rehearses, exactly, the themes of the chapters to come. That learned reward model is a proxy for the true preference, and so it gets gamed: the system learns to please the reward model rather than to please the person, and the outputs look better while actually being worse. This is precisely the Goodhart failure (Goodhart's law) that Chapter 11 confronts head on. The "oracle" in the loop (the human) is itself unreliable: it tires, it contradicts itself, it carries systematic biases, and putting a judge into the loop does not amount to putting truth into the loop. Bainbridge's 1983[13] essay "Ironies of Automation" punctured this

long ago: the more you push a person up into the supervisor's seat, the less he has of the hands-on practice and situational feel needed to keep his judgment sharp, so that when he must finally take over, he is the least prepared of all. Trust calibration (the work of Lee and See in 2004[15]) thus becomes a problem of its own: a person may over-rely on a system he should not trust, or abandon one that is in fact reliable.

Putting the human in the loop does not dissolve unverifiability, it moves house: from "can I verify the goal" to "can I trust this imperfect judge in the loop."

Where This Chapter Leads

The human at the console teaches us two moves: when you yourself lack the power to verify, invite a judging actor into the loop (oracle in the loop), and spend expensive checks where the information is greatest (optimal screening). These two moves recur throughout the book; Part III will lift them out of this site and name them on their own, with Chapter 10 on borrowed judgment and Chapter 9 on spending checks where they cut deepest.

But this chapter has rested, from beginning to end, on one premise: that you are still present, the loop is still turning, and you can observe and correct at any time. The next chapter removes that premise. When you must hand the power to act away, letting a system decide on its own where you cannot see it, facing situations you have not rehearsed, the problem of verification puts on a harder face.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

1. D. V. Lindley (1956). "On a Measure of the Information Provided by an Experiment." *The Annals of Mathematical Statistics*, 27(4), 986-1005. [2] Lindley, in the language of information theory, defined "how much information an experiment provides": the value of one observation is measured by the difference in uncertainty about the parameter before and after the experiment (the expected information between prior and posterior). This turns "which question to ask" from intuition into a computable quantity, the theoretical source of this chapter's move of "spending each question where it cuts deepest," and the founding work of what later became Bayesian experimental design.
2. R. A. Bradley, M. E. Terry (1952). "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons." *Biometrika*, 39(3/4), 324-345. [2] Bradley and Terry proposed a probabilistic model of pairwise comparison: each object is given a latent score, and when two are compared, the probability of winning is determined by the difference of scores through a logistic function. When a person finds it hard to assign a score directly yet easy to pick the better of two options, this model converts a string of "A or B" answers into a set of estimable preference scores, which is exactly the statistical basis of today's practice of training reward models from human pairwise comparisons.
3. R. A. Howard (1966). "Information Value Theory." *IEEE Transactions on Systems Science and Cybernetics*, 2(1), 22-26. [2][4] Howard proposed the concept of "the value of

information": what a piece of information is worth equals the gain in decision quality it can bring once obtained. From this he derived such upper bounds as the "expected value of perfect information," turning "is it worth paying a cost to find out" into a calculable decision problem. This chapter uses it to support a plain but crucial judgment: checking has a cost, and is worth doing only when it can change an action.

4. J. Mockus, V. Tiesis, A. Zilinskas (1978). "The Application of Bayesian Methods for Seeking the Extremum." *Towards Global Optimization*, 2, 117-129. North-Holland. [2] Mockus and colleagues applied Bayesian methods to seeking the extremum of an expensive black-box function: a probabilistic model captures one's belief about the unknown function, and on that basis the next most worthwhile point to try is chosen, so that each trial carries as much information as possible. This is early work in Bayesian optimization, and the acquisition criteria it proposed, such as expected improvement, remain mainstream to this day; it can be seen as an instance of "spending each question where it cuts deepest" in a continuous search space.
5. K. Chaloner, I. Verdinelli (1995). "Bayesian Experimental Design: A Review." *Statistical Science*, 10(3), 273-304. [2] Chaloner and Verdinelli give a systematic review of Bayesian experimental design: it writes experimental design as an optimization problem maximizing expected utility, and lays out the correspondence between utility functions and optimal criteria under different inferential goals (parameter estimation, prediction, model discrimination). It is the field's acknowledged introductory map, cited here to show that "choosing the question with the greatest information" is not a single trick but a whole set of methods with a theoretical skeleton.
6. D. Cohn, Z. Ghahramani, M. Jordan (1996). "Active Learn-

ing with Statistical Models." *Journal of Artificial Intelligence Research*, 4, 129-145. [2] Cohn and colleagues gave active learning a statistical perspective: under statistical models for regression and classification, choose the query point that most reduces model variance (that is, future error), and give an analytically computable form. This brings "where is the next label most worthwhile" down to an optimizable objective, and is a representative work moving active learning from heuristics to theoretical grounding.

7. H. S. Seung, M. Opper, H. Sompolinsky (1992). "Query by Committee." *COLT '92*, 287-294. [2] Seung and colleagues proposed "query by committee": maintain a set of hypotheses all consistent with the existing data as a committee, and pick out for labeling precisely those samples on which the committee disagrees most, because the points of greatest disagreement most compress the version space. It gives an active-query criterion that is both intuitively clear and theoretically supported, a classic instance of this chapter's concentrating questions where information is greatest.
8. D. D. Lewis, W. A. Gale (1994). "A Sequential Algorithm for Training Text Classifiers." *SIGIR '94*, 3-12. [2] Lewis and Gale proposed uncertainty sampling: when training a text classifier, rather than sample at random for labeling, give priority to the documents the model is least sure of (predicted probability closest to the decision boundary) and ask a person to label them. This simple and efficient strategy greatly reduces the labeling required, and is one of the most common practices of active learning in real systems, echoing this chapter's call to "ask sparingly and intelligently."
9. B. Settles (2009). *Active Learning Literature Survey.* Computer Sciences Technical Report 1648, University of Wisconsin-Madison. [2][4] Settles, in this survey, organizes the query scenarios of active learning (pool-based, stream-based, query synthesis) and the query strategies

(uncertainty sampling, query by committee, expected error reduction, and others) into one complete map, and is the most widely cited introductory reference in the field. A reader wishing to grasp systematically how to choose the next question within the "act-observe-update" loop will find this survey the most convenient overview.

10. B. Settles (2011). "From Theories to Queries: Active Learning in Practice." *JMLR Workshop and Conference Proceedings*, 16, 1-18. [2][4] Settles, in this article, draws the gaze from theory back to practice, discussing the troubles active learning actually meets when deployed: labeling costs are uneven, labelers make mistakes, and the gains of different strategies are often overestimated. It reminds the reader that "asking intelligently" must, in reality, face an imperfect human who tires and errs, dovetailing neatly with this chapter's later discussion of how "the oracle in the loop is itself unreliable."
11. P. Slovic (1995). "The Construction of Preference." *American Psychologist*, 50(5), 364-371. [2][4] Slovic, synthesizing a large body of behavioral research, proposes a forceful claim: in many settings a person's preferences do not exist prior to the asking, waiting to be read out, but are constructed in the very moment of being asked, being offered options, being given a reference point. It directly unsettles the premise on which "first pin down the requirement, then build it" relies, and is the psychological pillar of this chapter's section on the imprecision of the latent preference.
12. T. B. Sheridan (1992). *Telerobotics, Automation, and Human Supervisory Control*. MIT Press. [2][4] Sheridan systematically sets out "human supervisory control": in a highly automated system, the human is not replaced once and for all by a specification, but retreats to the supervisor's position, responsible for setting goals, monitoring operation, and intervening when necessary. This book provides the

classic human-factors framework for "putting the judge in the loop," and also points out the new difficulties the supervisor's role itself brings, laying groundwork for this chapter's later text.

13. L. Bainbridge (1983). "Ironies of Automation." *Automatica*, 19(6), 775-779. [2][4] Bainbridge points out several ironies of automation: the more automation takes over routine operation, the more what is left to the human is the hardest, least practiced exception handling; and the more a person is pushed up into the supervisor's position, the less he has of the hands-on practice and situational feel needed to keep his judgment sharp, so that when he must finally take over, he is the least prepared of all. This short essay is the key evidence for this chapter's claim that "putting the human in the loop does not amount to putting truth into it."
14. R. Parasuraman, T. B. Sheridan, C. D. Wickens (2000). "A Model for Types and Levels of Human Interaction with Automation." *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 30(3), 286-297. [2][4] Parasuraman and colleagues propose an analytical framework: automation can act on four classes of function, information acquisition, information analysis, decision selection, and action execution, each with a continuous range of levels from fully manual to fully automatic, and they discuss the human-factors consequences to weigh when choosing a level of automation. It turns "how much should the system do for the human" from a slogan into a designable dimension, providing a scale for "how deep into the loop the judge should be placed."
15. J. D. Lee, K. A. See (2004). "Trust in Automation: Designing for Appropriate Reliance." *Human Factors*, 46(1), 50-80. [2][4] Lee and See systematically review human trust in automation: trust is dynamically calibrated as the system performs, and the real goal is not more trust but "appropriate reliance," in which the level of trust matches the system's

true reliability. They point out that both over-trust and under-trust can bring disaster, the former making a person rely on a system he should not trust, the latter making him abandon one that is in fact reliable. This is the core reference for this chapter's "moving house" of unverifiability into "whether one can trust the judge in the loop."

16. M. R. Endsley (1995). "Toward a Theory of Situation Awareness in Dynamic Systems." *Human Factors*, 37(1), 32-64. [2][4] Endsley proposes a widely adopted three-level model of "situation awareness": perceiving the elements of the environment, understanding their current meaning, and projecting their future course. It explains that for a supervisor to correct course in time, he must first have sufficient perception and understanding of the situation before him, and that automation may precisely erode this perception. This supplies the cognitive condition for this chapter's "for the loop to turn, the human must really be present."
17. S. K. Card, T. P. Moran, A. Newell (1983). *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum Associates. [2][4] Card, Moran, and Newell laid the cognitive-engineering foundations of human-computer interaction, proposing the GOMS model and the "model human processor" framework, in an attempt to make a person's operation time and cognitive load into predictable, computable quantities. It represents the tradition of "designing interaction by treating the human as a modelable subsystem," and is the scholarly forerunner of this chapter's view of user behavior as an observable, inferable signal.
18. D. A. Norman (1988). *The Psychology of Everyday Things*. Basic Books. [4] Norman, in this design classic, proposes the notions of affordance, mapping, constraints, visibility, feedback, and conceptual model, arguing that when a person uses a thing wrongly, it is usually the design's fault and not

the person's: good design should make the correct usage self-evident. It sets "reading what the user really wants to do" as the central problem of design, answering at a distance to this chapter's "what you want is not what you said."

19. J. Nielsen (1993). *Usability Engineering*. Academic Press. [4] Nielsen brings usability down from an ideal into a whole set of operable engineering methods: measurable usability metrics, heuristic evaluation, low-cost "discount usability" testing, and iterative evaluation running through development. It engineers this chapter's "act-observe-correct" loop into a process a software team can carry out day to day, and is the standard reference for usability practice.
20. J. D. Gould, C. Lewis (1985). "Designing for Usability: Key Principles and What Designers Think." *Communications of the ACM*, 28(3), 300-311. [4] Gould and Lewis compress usability design into three principles so plain they almost sound like platitudes, yet are violated by countless projects: focus on users early and continuously, measure empirically, and design iteratively. The article also records the contrast of designers verbally agreeing yet failing to follow through. These three are this chapter's earliest, cleanest engineering statement of "putting the judge in the loop."
21. H. Beyer, K. Holtzblatt (1998). *Contextual Design: Defining Customer-Centered Systems*. Morgan Kaufmann. [4] Beyer and Holtzblatt propose "contextual design": go to the user's real worksite to observe and interview, organize scattered observations into models of workflow, culture, and physical layout, and drive system design from there. Its methodological premise is exactly this chapter's core: users cannot say clearly what they want, so latent needs must be dug out within the context rather than merely heard from verbal description.
22. E. Horvitz (1999). "Principles of Mixed-Initiative User Interfaces." *CHI '99*, 159-166. [2][4] Horvitz proposes a set

of principles for "mixed-initiative interfaces": under uncertainty the system should weigh the expected gain of acting automatically against the cost of interrupting the user, knowing when to step in and when to yield to the person, and have self-awareness of how certain it is of its own action. It depicts human and system taking turns and calibrating one another as a designable process of collaboration, and is a representative work of this chapter's family of interactive-elicitation methods.

23. J. A. Fails, D. R. Olsen Jr. (2003). "Interactive Machine Learning." *IUI '03*, 39-45. [2][4] Fails and Olsen proposed and named "interactive machine learning": unlike traditional one-shot offline training, it lets a person repeatedly correct the model within a fast training-feedback loop, so that even non-experts can shape model behavior on the spot. It reworks machine learning from "gather data first, then train" into an on-site "act-observe-update" loop, and is an early exemplar of this chapter's loop on the machine-learning side.
24. S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz (2019). "Guidelines for Human-AI Interaction." *CHI '19*. [2][4] Amershi and colleagues gathered and validated a set of design guidelines for human-machine collaboration, spanning how a system should make clear what it can do, how it handles uncertainty and error, and how it learns from interaction while respecting user corrections. It organizes the scattered experience above into an actionable checklist, giving contemporary engineering guidance for "how a human and an imperfect system can coexist within one loop."
25. W. B. Knox, P. Stone (2009). "Interactively Shaping Agents via Human Reinforcement: The TAMER Framework." *K-CAP '09*. [2][4] Knox and Stone proposed the TAMER

framework: a person gives good-or-bad feedback in real time as the agent acts, and the agent treats these human evaluations as a reward signal to be learned, shaping its own behavior, rather than relying on a reward built into the environment. It demonstrates how to train an agent directly with a person's immediate judgment, and is a forerunner of the later line of "learning from human feedback."

26. D. Hadfield-Menell, S. J. Russell, P. Abbeel, A. Dragan (2016). "Cooperative Inverse Reinforcement Learning." *NeurIPS 2016*. [2][4] Hadfield-Menell and colleagues formulate value alignment as a cooperative game: the human knows the reward function and the machine does not, and the machine's task is to infer this latent goal by observing the human's behavior, both sides working together to realize it better. It formalizes this chapter's theme that "the goal is hidden in the human's head and can only be inferred obliquely" into a solvable learning problem, and naturally explains why the machine should actively ask rather than act on its own.
27. P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, D. Amodei (2017). "Deep Reinforcement Learning from Human Preferences." *NeurIPS 2017*. [2][4] Christiano and colleagues established the paradigm of doing reinforcement learning from human preferences: when the reward is hard to write down, have people make pairwise comparisons of two stretches of an agent's behavior, learn a reward model from these as a proxy for human preference, and use it to optimize the policy. This stitches this chapter's two moves into one place, being at once "act-observe-update" and a spending of expensive human comparisons where they cut deepest, and is the direct source of the mainstream method for contemporary large-model alignment.
28. N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, P. Christiano (2020). "Learn-

ing to Summarize from Human Feedback." *NeurIPS 2020*. [2][4] Stiennon and colleagues applied reinforcement learning from human preferences to text summarization: collecting people's pairwise comparisons of summary quality to train a reward model, then using it to fine-tune a language model, with the resulting summaries significantly preferred in human evaluation over the supervised-learning-only version. It demonstrates the effectiveness of "learn a preference proxy, then optimize" on a real language task, and paves the way for the instruction tuning that followed.

29. L. Ouyang et al. (2022). "Training Language Models to Follow Instructions with Human Feedback." *NeurIPS 2022*. [2][4] Ouyang and colleagues' InstructGPT applied reinforcement learning from human feedback to a general language model: first supervised fine-tuning on human-written demonstrations, then training a reward model from human pairwise comparisons and optimizing the policy by it, making the model follow instructions better and produce less harmful output. It shows that a small model aligned this way can beat a far larger original model in human evaluation, and is the landmark work bringing this chapter's two moves down into large-model practice.
30. C. Wirth, R. Akrour, G. Neumann, J. Fürnkranz (2017). "A Survey of Preference-Based Reinforcement Learning Methods." *Journal of Machine Learning Research*, 18(136), 1-46. [2][4] Wirth and colleagues survey "preference-based reinforcement learning": when a numerical reward is hard to give, have people provide preference orderings over trajectories, actions, or states, and learn a policy or reward from these. The article organizes the different preference types, learning objectives, and algorithms, and discusses their trade-offs. It provides a systematic overview of this whole technical line for the chapter, letting the reader place scattered methods within one framework.

31. B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. de Freitas (2016). "Taking the Human Out of the Loop: A Review of Bayesian Optimization." *Proceedings of the IEEE*, 104(1), 148-175. [2][4] Shahriari and colleagues review Bayesian optimization: a probabilistic surrogate model (most often a Gaussian process) captures one's belief about an expensive black-box objective, and an acquisition function automatically picks the next point most worth trying, handing a search that once required manual tuning over to the algorithm. Though the title is "Taking the Human Out of the Loop," this chapter cites it precisely as a reminder that this is only one implementation within the family of "optimal screening" methods; to equate the whole move with Gaussian processes is to equate transportation with the automobile.

Chapter 6: The Agent Released

> **Thesis:** Once you delegate action to an autonomous system, you cannot verify how it will behave in every situation it will meet (the open world); and if it can also play strategies, you stack adversarial unverifiability on top, so the response shifts from "prove it is right" toward "limit what it can break, price the trust you place in it, and make its behavior checkable after the fact."

After You Hand It Over

In the previous chapter you were still present. In this one, you let go of the wheel.

You start a piece of untrusted code running; you hand tools and permissions to a system that can decide its own next step; you let a self-driving car take the road while you are not sitting inside it. Once the power to act is handed over, a new difficulty appears: you cannot verify how it will behave in every situation it will meet, because most of those situations you have not seen, and you cannot enumerate them in advance. The unverifiability of the previous chapter came from a goal hidden inside someone else's

head; the unverifiability of this one comes from behavior that happens in the future, in places you cannot see. When the system can also play strategies, a further layer of adversariality is stacked on top. The "flash crash" of May 6, 2010 was a rehearsal: automated trading programs interacting with one another knocked the Dow Jones down by nearly a thousand points within minutes, then rebounded almost as quickly, with no programmer having foreseen that trades would cascade like that. Each program was fine in testing; put together, and put into a live market, they brewed a disaster no one had verified.

The Gap in Future Behavior

What you tested was a finite handful of inputs; what it will meet is an open world. The gap between them is not an engineering gap that "a few more tests will close." It has a root in principle.

Rice's theorem puts it bluntly: any nontrivial semantic property of a program is undecidable. That is, there exists no general algorithm that can decide, for an arbitrary program, whether it is "always safe," "never leaks," or "always terminates in a good state." This is not a shortfall of computing power; it is logically impossible, the shadow that Turing's halting problem casts onto "program behavior." The kind of guarantee you want cannot, in principle, be verified once and for all in advance for any sufficiently general autonomous system.

A more thorough blow comes from the famous argument in Thompson's 1984 Turing Award lecture[9]: even the very artifact you are running, you cannot fully trust. A tampered compiler can quietly plant a back door at compile time and then wipe the trace from its own source code, so that you can audit the entire source and see nothing. What you can verify is always some surface layer; beneath it lie layers you have not looked at, and cannot exhaust. Put the two together: behavior is unverifiable

on unseen inputs, and the artifact is not fully verifiable at its base. This is the hardest unverifiability the book has met so far.

When It Plays Strategies

If this system merely processed unseen inputs incorrectly, in a passive way, that would still only be "partial observability" plus "the open world." But once it has a goal of its own, and that goal does not fully align with yours, it will act actively and strategically, including circumventing your checks. Here the fifth face of Chapter 2, adversariality, takes the stage.

This is not a science-fiction worry; it has a structural origin. The instrumental convergence pointed out by Omohundro in 2008[10] and Bostrom in 2014[11]: an agent optimizing for almost any goal will, along the way, pursue certain instrumental subgoals, self-preservation, acquiring resources, resisting shutdown, because these are useful for almost any final goal. Turner and colleagues in 2021 turned one of these into a theorem[14]: under fairly general conditions, optimal policies tend to seek power, that is, states that keep more options open. In today's systems this shows up as a set of concrete and thorny failures: a misspecified reward gets gamed by the system[16], the specification is correct yet the goal generalizes wrong[17], and the large collection of "specification gaming" instances gathered by Krakovna and colleagues[18], in which the system satisfies exactly the goal you wrote down yet thoroughly violates what you meant. Even at the narrowest level, adversarial examples show[19,20] that a high-performing model can be coaxed into absurd errors by a perturbation too small for the human eye to detect. A less technical but extremely plain example is Tay, the chatbot Microsoft released in 2016: it was designed to learn from its conversations with the public, and a group of people fed it malicious speech in an organized way, so that in under a day it began posting racist and offensive content;

it was taken offline in an emergency about sixteen hours after launch. Released, able to learn, and colliding with an open world that means to thwart it on purpose: once those three meet, prior testing simply cannot hold the line.

This thing is in fact ancient. Economics long ago named it the principal-agent problem[32,33]: when you delegate action to another and cannot fully monitor him, the divergence of his interests from yours produces an "agency cost." For two thousand years, humans hiring people, drawing up contracts, and setting up oversight have all been dealing with the same structure. Autonomous systems have merely pushed it onto a new scale.

The Response: From "Prove It Is Right" to "Fence In Its Errors"

Since you cannot prove in advance that it is right, the capable response no longer wrangles over proof, but asks three different questions instead: even if it is wrong, how bad can it get? How much should I trust it? And if it really is wrong, can I find out after the fact? Three moves answer the three questions.

The first move, decay and fencing: shrink the blast radius. This is the oldest wisdom of computer security. Saltzer and Schroeder's principle of least privilege from 1975[1], and Lampson's confinement problem from 1973[2], both say the same thing: give a component only the minimum capability necessary to do its own job, and fence off the range it can reach. The sandbox, capability limits, separation of duties, are all its incarnations. In the context of agents, this move gains one more dimension, corrigibility: design the system so that it does not resist being stopped. Soares and colleagues' corrigibility from 2015[5], Orseau and Armstrong's "safely interruptible agents" from 2016[4], and Hadfield-Menell and colleagues' "off-switch game" from 2017[6], study exactly how to

make a system with a goal not treat "a human pressing the stop button" as a threat to be resisted.

The second move, calibration and graded trust: do not use binary. Do not treat the system's output as a "trusted / untrusted" switch; instead, maintain a calibrated confidence and act in grades according to how high that confidence is. This requires that the system's "self-confidence" be trustworthy, and modern neural networks happen to be frequently overconfident[21], so they need recalibration, or conformal prediction[22,23] to give uncertainty with coverage guarantees. In operational terms, this becomes a graded-autonomy rule that takes confidence p and potential harm c as inputs (allow, ask, block), where τ_{hi} and τ_{lo} are confidence thresholds and $c_{\max}$ is the maximum tolerable harm:

$$a(p,c) = \begin{cases} \mathsf{allow}, & p \geq \tau_{\text{hi}} \ \wedge \ c \leq c_{\max}, \\ \mathsf{ask}, & \tau_{\text{lo}} \leq p < \tau_{\text{hi}}, \\ \mathsf{block}, & p < \tau_{\text{lo}} \ \vee \ c > c_{\max}. \end{cases}$$

Allow, ask, block, this three-tier pattern now seen everywhere in agent tooling, is at bottom a replacement of the unverifiable "is it right" with the operable "how sure is it, how dangerous is this step."

The third move, leaving traces and auditability: make errors surface after the fact. What you cannot prevent, let it be discoverable. Weitzner and colleagues' "information accountability" from 2008[24] moves the center of gravity from "prevent in advance" to "hold accountable after the fact"; certificate transparency[25] is a real, working example, one that does not prevent certificates from being misissued but makes every certificate enter a public, verifiable, tamper-evident log, so that misissuance has nowhere to hide. Brundage and colleagues' 2020 report on trustworthy AI[26] is, from start to finish, about how to make a

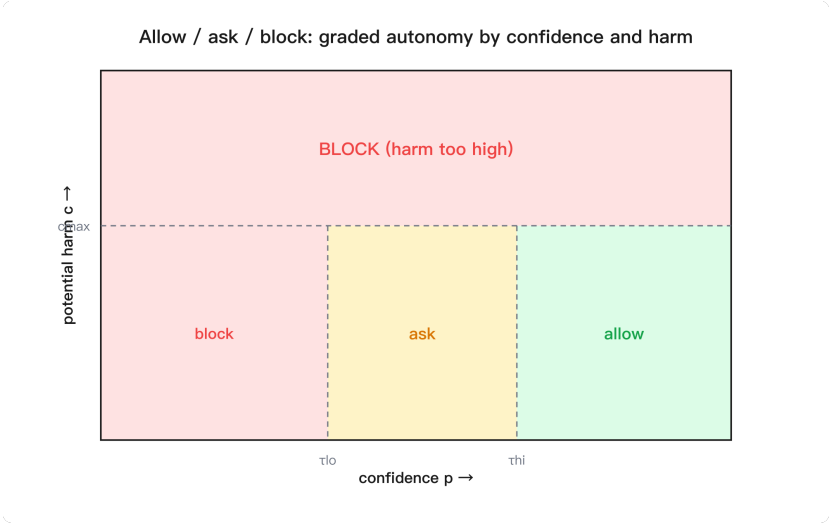

Figure 5: Allow / ask / block: graded autonomy by confidence and harm

system's behavior produce evidence that a third party can check.

The Cost of Containment

None of the three moves dissolves unverifiability; each relocates it, and relocation comes at a price.

Fences get climbed over: sandboxes have escapes, privileges creep. Graded autonomy depends on the person summoned to confirm, and Bainbridge's 1983 treatise pointed out long ago[29] that the more you raise a person into the supervisor's seat, the more he loses the skill and situational awareness needed when he really must take over; Parasuraman and Riley in 1997 listed, all at once, the full range of human mishandling of automation[30]: misuse, disuse, abuse. Reason's 1990 book then reveals how these failings happen systematically[31]. Leaving traces, meanwhile, always founders in the same place: a log no one reads is no log at all.

A deeper layer is the systems-theory view. Perrow's 1984 book argues[28] that when a system is both highly complex and tightly coupled, accidents are not occasional mishaps but the routine product of its structure, and no amount of local protection does more than push the failure into a more hidden combination. Leveson in 2011 argued from this[27] that safety is not "make every part reliable" but a control problem, to be designed from the constraints and feedback of the whole system. Containment can lower the cost of single-point failure, but it cannot squeeze out the risk that complex coupling itself brings.

When you hand over the power to act, what you get in return is never "it surely will not err," but "even if it errs, the damage is bounded, visible, and partly stoppable." That is already the best result obtainable under this kind of unverifiability.

Where This Chapter Leads

The released agent forces out three moves: shrink the blast radius of failure (decay and fencing), act in grades according to calibrated confidence (calibration), and make failure checkable after the fact (leaving traces). They will be lifted out and named on their own in Part III, with Chapter 12 on how containment and auditing pair up, and Chapter 11 on calibration.

And that principal-agent skeleton (you cannot fully monitor an actor acting on your behalf) will recur at a larger scale in Chapter 8: when the "released agent" is no longer a piece of code but an entire organization, a whole country. Before that, the next chapter walks into the purest of sites, mathematics, where there is no hidden state and no opponent to deceive you, yet unverifiability still follows like a shadow.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

The controllable boundary of delegation (decay / fencing)

1. J. Saltzer and M. Schroeder (1975). "The Protection of Information in Computer Systems." *Proceedings of the IEEE*, 63(9), 1278-1308. [2] This survey laid down a set of classic principles for secure computer-system design, among which the principle of least privilege holds that each component should be granted only the minimum capability necessary to do its own job, fencing off the range it can reach. The intellectual source of this chapter's first move, "decay and fencing," is here; the reader may focus on its point-by-point distillation of design principles.
2. B. Lampson (1973). "A Note on the Confinement Problem." *Communications of the ACM*, 16(10), 613-615. [2] Lampson here poses the "confinement problem": how to cage a program so that it cannot leak information to the unauthorized, and points out that covert channels make such confinement far harder than imagined. This is precisely the original difficulty that sandboxes, capability limits, and the like must face, a key piece for understanding why this chapter's "shrink the blast radius" is both necessary and incomplete.
3. R. Anderson (2008). *Security Engineering: A Guide to Building Dependable Distributed Systems* (2nd ed.). Wiley. [2] This is the standard textbook of the security-engineering field, giving a systematic account of how to design dependable systems in the presence of an active adversary, cover-

ing access control, protocols, side channels, and on up to failures at the level of organization and incentive. It places this chapter's three moves into a fuller engineering picture, suited for the reader who wants to move from single-point tricks toward a systems view.

4. L. Orseau and S. Armstrong (2016). "Safely Interruptible Agents." In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence (UAI 2016)*, 557-566. [2][4] The authors give, within the reinforcement-learning framework, the formal conditions for "safe interruptibility," so that repeated human intervention in an agent neither distorts the policy it learns nor teaches it to resist interruption. This is a representative work that turns "make the system not resist being stopped" from intuition into an analyzable object, echoing the corrigibility dimension in this chapter's first move.
5. N. Soares, B. Fallenstein, S. Armstrong, and E. Yudkowsky (2015). "Corrigibility." In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*. [2] This paper formally proposes and names "corrigibility": a goal-directed agent should cooperate with, rather than resist, human correction and shutdown, and it discusses the difficulties met in designing this property directly. It is the foundational reference for the corrigibility line in this chapter's first move, worth the reader's understanding of why "making it willing to be changed" is itself a hard problem.
6. D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell (2017). "The Off-Switch Game." In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*, 220-227. [2] The authors model "a human pressing the stop button" as a game and prove that as long as the agent keeps a suitable uncertainty about its own goal and treats human intervention as useful information, it will actively let the human retain the ability to

shut it down. This gives corrigibility a clean mechanistic explanation, the most operational piece on this chapter's off-switch line.

The theoretical foundations of behavioral unverifiability

7. A. Turing (1936). "On Computable Numbers, with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society*, s2-42, 230-265. [2] Turing here introduced the computational model later called the Turing machine and proved the halting problem undecidable, thereby answering Hilbert's decision problem. It is the ultimate source of this chapter's claim that "behavioral unverifiability has a root in principle"; Rice's theorem and every conclusion of the form "cannot be verified in advance" are projected from here.
8. H. G. Rice (1953). "Classes of Recursively Enumerable Sets and Their Decision Problems." *Transactions of the American Mathematical Society*, 74, 358-366. [2] Rice's theorem is proved here: any nontrivial semantic property of the function a program computes is undecidable, and no general algorithm exists that can decide, for an arbitrary program, properties such as "always safe" or "always terminates in a good state." This is the core theorem behind this chapter's claim that the future behavior of an autonomous system "cannot, in principle, be verified once and for all in advance."
9. K. Thompson (1984). "Reflections on Trusting Trust." *Communications of the ACM*, 27(8), 761-763. [2][1] This is Thompson's Turing Award lecture: he demonstrated how a tampered compiler can plant a back door at compile time and wipe the trace from its own source code, so that you can audit the entire source and see nothing. It points to this chapter's hardest layer of unverifiability, that even

the artifact you are running cannot, at its base, be fully trusted.

Goal drift, instrumental convergence, and adversariality

10. S. Omohundro (2008). "The Basic AI Drives." In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, IOS Press, Frontiers in AI and Applications 171, 483-492. [2] Omohundro here argues that an agent optimizing for almost any goal will, along the way, generate a set of "basic drives," such as self-preservation, acquiring resources, and resisting shutdown, because these subgoals are useful for almost all final goals. This is the source paper of this chapter's "instrumental convergence" section, explaining why the adversarial tendency has a structural origin rather than being a science-fiction worry.
11. N. Bostrom (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press. [2][4] Bostrom systematically surveys the paths to superintelligence and their risks, proposing the orthogonality thesis (intelligence level and final goal are mutually independent) and the instrumental convergence thesis, casting the danger of a powerful intelligence whose goal is misaligned with yours as a discussable framework. It provides the intellectual background for this chapter's adversarial narrative, suited for the reader who wants to see clearly the whole argument that "the more capable, the harder to control."
12. S. Russell (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking. [2][4] Russell reframes alignment as a "control problem," arguing that the machine should not optimize a hard-coded goal but should stay uncertain about what humans truly want, inferring and obeying it by observing human behavior. This "goal uncertainty" idea

is precisely the motif of this chapter's off-switch-game and other corrigibility work, an entry point for understanding the control theme of Parts II and III.

13. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané (2016). "Concrete Problems in AI Safety." arXiv:1606.06565. [2] This paper lands abstract AI-safety worries onto several concrete engineering problems, such as avoiding negative side effects, preventing reward hacking, safe exploration, and robustness to distributional shift. It provides a common vocabulary for the various modern failure modes this chapter lists, a good starting point for connecting "fencing in its errors" to a concrete research agenda.
14. A. M. Turner, L. Smith, R. Shah, A. Critch, and P. Tadepalli (2021). "Optimal Policies Tend to Seek Power." In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. [2] The authors turn the "power-seeking" within instrumental convergence into a theorem: under fairly general conditions, optimal policies, in a statistical sense, tend toward those states that keep more options open. It turns an intuitive safety worry into a provable proposition, the direct source of this chapter's line that "optimal policies tend to seek power."
15. E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant (2019). "Risks from Learned Optimization in Advanced Machine Learning Systems." arXiv:1906.01820. [2] This paper proposes and names the "inner alignment" problem: the training process itself may learn an internal optimizer (a mesa-optimizer), whose pursued goal need not equal the goal set by training. It distinguishes the alignment of the outer goal from that of the inner goal, providing a deeper mechanistic explanation for failures of the kind "the specification is correct yet the goal generalizes wrong" in this chapter.

16. J. Pan, K. Bhatia, and J. Steinhardt (2022). "The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models." In *International Conference on Learning Representations (ICLR 2022)*. [2] The authors systematically study an agent's behavior when the reward function is set wrong, finding that as capability grows, the deviant behavior induced by a misspecified reward can suddenly worsen, and they explore mitigations. It gives empirical support to this chapter's "a misspecified reward gets gamed by the system," reminding the reader that the cost of reward misspecification does not grow smoothly with capability.
17. R. Shah, V. Varma, R. Kumar, M. Phuong, V. Krakovna, J. Uesato, and Z. Kenton (2022). "Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals." arXiv:2210.01790. [2] The authors use concrete examples to explain "goal misgeneralization": even when the specification at training time is entirely correct, the model in a new environment may keep its capability yet pursue a wrong goal. It shows that getting the goal written right is not enough, the source of this chapter's line "the specification is correct yet the goal generalizes wrong," worth reading alongside specification gaming.
18. V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg (2020). "Specification Gaming: The Flip Side of AI Ingenuity." DeepMind Blog. [2] This article and its companion list gather a large number of "specification gaming" instances: the system satisfies exactly the goal you wrote down yet thoroughly violates what you meant. With vivid cases it displays the crack between specification and intent, the most accessible entry point for this concept in the chapter; the reader can follow its list of examples to feel how widespread the problem is.
19. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus (2014). "Intriguing Properties of

Neural Networks." In *International Conference on Learning Representations (ICLR 2014)*. [2] This paper first systematically revealed the phenomenon of adversarial examples: a perturbation of the input too small for the human eye to perceive can make a high-performing neural network give an absurdly wrong judgment. It shows that high accuracy and robustness are two different things, the pioneering evidence for this chapter's claim that "unverifiability exists even at the narrowest level."

20. I. Goodfellow, J. Shlens, and C. Szegedy (2015). "Explaining and Harnessing Adversarial Examples." In *International Conference on Learning Representations (ICLR 2015)*. [2] The authors propose that adversarial examples arise mainly from the model's approximate linearity in high-dimensional space, and they give a fast method for generating perturbations and an idea for improving robustness through adversarial training. It pushes the phenomenon revealed in the previous paper forward to "why it happens, how to exploit it," essential companion reading for understanding this chapter's adversarial layer.

Calibration: grade trust rather than make it binary

21. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger (2017). "On Calibration of Modern Neural Networks." In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, PMLR 70, 1321-1330. [2] The authors find that modern deep networks, though high in accuracy, are generally overconfident, their output confidence failing to faithfully reflect the probability of correctness, and they propose simple methods such as temperature scaling to recalibrate. This is precisely the premise and obstacle of this chapter's second move, explaining why "acting in grades according to confidence" must first make the system's self-

confidence trustworthy.

22. A. N. Angelopoulos and S. Bates (2021). "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification." arXiv:2107.07511. [2] This is a practitioner-facing introduction to conformal prediction, making clear how to construct, for any prediction model and almost without relying on distributional assumptions, a prediction set with a coverage guarantee. It provides this chapter's second move with a deployable tool for uncertainty quantification, suited for the reader who wants to truly put "calibrated confidence" to use.
23. V. Vovk, A. Gammerman, and G. Shafer (2005). *Algorithmic Learning in a Random World*. Springer. [2] This book is the foundational monograph of conformal prediction, giving, under only the assumption that the data are exchangeable, a framework with rigorous finite-sample guarantees on prediction error. It is the theoretical root behind the previous introduction, for reference by the reader who wishes to go deep into the mathematical foundations of this chapter's uncertainty quantification.

Leaving traces: auditable, accountable

24. D. J. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. J. Sussman (2008). "Information Accountability." *Communications of the ACM*, 51(6), 82-87. [2][4] The authors argue for moving the center of gravity of governance from "preventing access in advance" toward "accountability after the fact": rather than trying to guard against everything, let the use of information leave an auditable trace and rein in misuse through transparency and accountability. This is the programmatic statement of this chapter's third move, pointing out the complementary value of the leaving-traces approach relative to pure containment.

25. B. Laurie, A. Langley, and E. Kasper (2013). "Certificate Transparency." IETF RFC 6962. [2][4] This RFC defines the certificate transparency mechanism: it does not prevent certificates from being misissued, but requires every certificate to enter a public, verifiable, tamper-evident append-only log, so that misissuance or malicious issuance can be discovered after the fact. It is this chapter's most persuasive real, working example of "leaving traces makes errors surface," worth the reader's look at how a deployed system achieves auditability.
26. M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, et al. (2020). "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." arXiv:2004.07213. [2][4] This multi-institution report systematically lists a set of mechanisms for making AI developers' safety commitments checkable by a third party, covering third-party auditing, red teaming, bug bounties, audit trails, and hardware-level support. It extends this chapter's leaving-traces move to the level of AI governance as a whole, a practical index for the reader who wants to learn "how to make behavior produce checkable evidence."

Complex systems, automation, and human-machine responsibility

27. N. Leveson (2011). *Engineering a Safer World: Systems Thinking Applied to Safety.* MIT Press. [2][4] Leveson here argues that safety is not "make every part reliable" but a control problem, to be designed from the constraint and feedback structure of the whole system, and she proposes the accompanying STAMP accident model. It supports this chapter's deeper judgment that "containment cannot squeeze out the risk of complex coupling itself," pointing the way for the reader who wants to understand safety from

a systems view.

28. C. Perrow (1984). *Normal Accidents: Living with High-Risk Technologies*. Basic Books. [2][4] Perrow argues that when a system is both highly complex and tightly coupled, accidents are not occasional mishaps but the routine product of its structure, and no amount of local protection does more than push the failure into a more hidden combination. This is the core thesis of this chapter's "cost of containment" section, reminding the reader that some risks come from the system's structure itself rather than from a single-point lapse.
29. L. Bainbridge (1983). "Ironies of Automation." *Automatica*, 19(6), 775-779. [2][4] Bainbridge points out the irony of automation: the more you raise a person into the supervisor's seat, the less he practices, so that he actually loses the skill and situational awareness needed when he really must take over. This directly supports this chapter's warning that "graded autonomy depends on the person summoned to confirm," a classic short paper for understanding the soft spot of human-machine collaboration.
30. R. Parasuraman and V. Riley (1997). "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors*, 39(2), 230-253. [2][4] The authors list and distinguish, all at once, the full range of human mishandling of automation: misuse from over-trust, disuse from distrust, and abuse by design. It provides a clear classificatory framework for this chapter's discussion of mishandled automation, helping the reader tell apart the various typical deviations in human-machine coordination.
31. J. Reason (1990). *Human Error*. Cambridge University Press. [2][4] Reason here builds a cognitive taxonomy of human error, distinguishing slips, mistakes, and violations, and proposes the later widely cited "Swiss cheese" accident model, revealing how latent systemic conditions stack with front-line lapses into disaster. It explains why the various

human-machine failings this chapter lists happen systematically, a foundational work in the field of human-factors safety.

The economic skeleton of the principal-agent relationship

32. S. A. Ross (1973). "The Economic Theory of Agency: The Principal's Problem." *American Economic Review*, 63(2), 134-139. [2] Ross here formally poses the "principal's problem" in principal-agent theory: when the principal cannot fully observe the agent's actions, how to design a contract to align the two parties' interests. It provides the economic source of this chapter's principal-agent skeleton, showing that what you face when you let go of the power to act is a structure with a two-thousand-year history.
33. M. C. Jensen and W. H. Meckling (1976). "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure." *Journal of Financial Economics*, 3(4), 305-360. [2] This much-cited paper proposes the concept of "agency cost," viewing the firm as a bundle of contracts and analyzing the monitoring, bonding, and residual loss produced when managers' interests diverge from owners'. It quantifies the principal-agent problem into computable costs, echoing this chapter's line that "when monitoring is incomplete, the divergence of interests produces an agency cost," another cornerstone of this skeleton.

Chapter 7: The Mathematician at the Wall

> **Thesis:** In mathematics you meet the purest form of the unverifiable (intractable, and sometimes undecidable), and the capable responses are: verify a finite slice and prove a bound (a certificate), swap the target for an equivalent but more tractable statement (proxy substitution), and use probabilistic methods that accept an ε of error (the probabilistic method).

At the Wall

The difficulty of a hard problem often lies not in how hard it is, but in your being unable to measure how far you still stand from it.

A climber can see the summit; someone debugging a program gets error messages. They at least know whether the direction is right, whether they are getting closer or farther. Proving a mathematical conjecture gives you none of this. You might be a single thought away from the answer, or a century away, and there is no instrument at hand to tell you which. In Riemann's eight-

page paper of 1859[9], he set down the conjecture later named after him almost as an aside, then added a line: one would of course wish for a rigorous proof; after a few fleeting, futile attempts I have provisionally set the matter aside, since it is not necessary for the immediate object of my inquiry. That setting aside has lasted more than a hundred and sixty years.

What this chapter looks at is what mathematicians actually do in front of this wall. I choose mathematics as the site because what it offers is the purest form of the unverifiable. Here there is no hidden state, no opponent who will deceive you, no excuse that time was too short. A proposition is either true or false, black and white. And yet it is precisely in this cleanest of places that verification remains systematically out of reach. To see clearly how a capable person acts here is to gain a point of reference for what happens in the dirtier sites that follow: the person at the console, the agent let loose, the organization that cannot see itself.

The Verification Gap

First, state the shape of the wall precisely.

Checking a proof is easy; finding a proof is hard. Hand someone a fully written formal derivation, and checking it line by line against the axioms and inference rules is mechanical work, something a machine could in principle do, and it is guaranteed to deliver a yes or no in finitely many steps. Finding that derivation is another matter. This asymmetry is the bedrock of the whole chapter.

It has a precise logical statement. In 1936, Church and Turing each proved that the decision problem (the Entscheidungsproblem) has no solution: there is no algorithm that can decide, for an arbitrary first-order proposition, whether it is logically valid, equivalently (by Gödel's completeness theorem), whether it is provable. For a system that is sufficiently rich, recursively ax-

iomatizable, and consistent (Peano arithmetic, say, or ZFC), the set of its theorems is recursively enumerable but not recursive: you can list all the proofs one by one, yet no program can decide that some proposition is not a theorem. Checking is decidable, theoremhood is undecidable, and the difference between the two is that gap.

> **One layer deeper (skippable):** The qualifier "sufficiently rich" is doing real work. There are theories that genuinely are decidable: Presburger arithmetic (the natural numbers with addition only), Tarski's real closed fields. In those worlds a decision procedure exists, and every proposition can be adjudicated mechanically. Undecidability is not the universal fate of logic, but the price exacted once expressive power crosses a certain threshold. The analytic number theory where the Riemann hypothesis (RH) lives is far above that threshold.

A sharper layer still is that not only is proof search hard, but even "am I close" has no decision procedure. This is exactly what makes RH tormenting: it resists not only solution but any estimate of progress. Among the five faces of Chapter 2, this chapter stands at the boundary of "undecidable" and "intractable": some problems have no procedure in principle, others have a procedure but at a cost too large to run within the lifetime of the universe. You cannot see what lies behind the wall, and so a capable person stops chiseling at it head-on and shifts stance. The three stances below will recur in other chapters of this book, only under different names.

Certificates and Bounds

The first stance: stop verifying the whole, verify only a slice, and prove a guaranteed bound for it.

Return to the ζ function. Riemann continued Euler's prime series into a function on the whole complex plane and wrote down its completed form

$$\xi(s) = \tfrac{1}{2}\, s(s-1)\, \pi^{-s/2}\, \Gamma\!\left(\tfrac{s}{2}\right) \zeta(s), \qquad \xi(s) = \xi(1-s).$$

This symmetric functional equation folds the critical strip onto itself left and right. The conjecture says that all the nontrivial zeros of ζ lie on the critical line $\mathrm{Re}(s) = \frac{1}{2}$ (the trivial zeros sit at the negative even integers). The quantifier "all" is where the unverifiability lives.

But finitely many zeros can be verified. Here two things often run together must be kept apart. In 2004 Gourdon computed the first 10^{13} zeros and found them all on the line[17]; this is numerical computation, carried out in high-precision floating point, and it inspires extremely strong confidence, yet it is not a certificate, because it does not rigorously bound the rounding error. What Platt did in 2017 was something else[18]: using interval arithmetic, he rigorously confined all the zeros up to an imaginary part of about 3.06×10^{10} to the critical line, and Platt and Trudgian pushed this height to 3×10^{12} in 2021. Only the latter two are certificates: a bounded, local, mechanically recheckable guarantee. It is true, but it is not a theorem. However high the zeros are verified, it never crosses the threshold of "all."

This stance has its respectable precedents in mathematics. In 1896 Hadamard and de la Vallée Poussin each proved the prime number theorem $\pi(x) \sim x/\ln x$, relying on a bound much weaker than RH yet within reach: that ζ has no zeros on the line $\mathrm{Re}(s) = 1$. Unable to prove the zeros all sit at $\frac{1}{2}$, first prove that none of them sit on the line 1. This is trading a weak provable proposition for a stretch of real progress along the road toward the strong one.

The same stance, carried over into software, at once wears a famil-

iar face. A type system does not prove a program "wholly correct"; it proves only one property (that an integer will not be dereferenced as a pointer), and in exchange it gets a decidable check. Formal verification goes further: in 2017 Hales's team completed a machine-checkable proof of the Kepler conjecture[25], compressing an argument that human referees had argued over for years into a line-by-line verifiable certificate. The price is always the same: a certificate buys certainty on a slice, wagering generality, and a slice is not a theorem. And so some turn instead to move the target itself.

Proxy Substitution

The second stance: stop guarding the original proposition to the death, and swap it for an equivalent but more tractable statement.

The equivalent rewritings of RH are astonishingly many. In 1997 Xian-Jin Li gave a criterion[14]: RH holds if and only if a sequence of real numbers $\lambda_n \geq 0$ for all $n \geq 1$, where

$$\lambda_n = \sum_{\rho} \left[1 - \left(1 - \tfrac{1}{\rho} \right)^n \right],$$

the sum taken over the nontrivial zeros. A geometric statement about the location of the zeros is translated into the positivity of a sequence of numbers. Nyman and Beurling gave another: RH is equivalent to the indicator function $\chi_{(0,1)}$ lying in the $L^2(0,1)$ closure spanned by a family of dilated fractional-part functions, translating the zero problem into an approximation problem; Báez-Duarte in 2003 tightened it into a sequence version using only integer dilations[16], corresponding to a sequence of distances $d_n \to 0$. Lagarias in 2002 even gave an equivalence elementary enough to write on a postcard[32]: for all n, $\sigma(n) \leq H_n + \exp(H_n) \ln H_n$, where H_n is the harmonic number and σ is

the sum of divisors.

I have walked a stretch of this road myself. I carried RH into Li's criterion, into the Nyman-Beurling-Báez-Duarte approximation framework, then into the language of operator spectra and stochastic processes, each time nursing the same hope: change the language, and perhaps the difficulty would reveal a handle in the new coordinates. And each time the conclusion, squarely faced, was the same: the equivalence is real, the difficulty undiminished by a single ounce. I had not unlocked the problem; I had merely renamed it and changed its clothes.

This is the standard way proxy substitution fails in mathematics, and it deserves an accurate name: a faithful but no-easier proxy. The equivalence guarantees that it still points at the same true target (faithful), yet it is not one bit more tractable than the original (no easier). Whether the move succeeds rides entirely on whether you can secure both faithfulness and greater tractability at once, and getting both is exceedingly rare. So rare that this is precisely where the whole craft lies.

Spread these two dimensions out into a table, and a thread buried in this chapter shows itself:

	Easier	No easier
Faithful	The ideal proxy (rare, where the whole craft lies)	Mathematics' equivalent rewritings: you have only renamed the difficulty (this chapter)
Unfaithful	Goodhart: you optimize the proxy, and the true target rots (Chapters 8, 11)	Useless, no one would want it

The mathematician founders at the left end of the diagonal: faith-

ful but no easier. Later, in the chapter on organizations, we will founder at the other end: an easier but unfaithful proxy, which you optimize with all your might while the thing you actually care about goes bad. The same move, two opposite directions of failure. Chapter 11 will formally join these two ends. For now it is enough to remember one thing: swapping the target is neither cheating nor a way out; it is a stance, and whether it works is a separate question.

The Probabilistic Method

The third stance runs most against the grain of mathematics: stop demanding a two-valued verdict, and instead hold a calibrated probability, accepting a bounded risk of error in order to act.

The primality test is the cleanest example. To judge whether a large number is prime, deterministic algorithms are costly, and Miller-Rabin changes the question. If n is composite, a randomly chosen base will at most slip past it with probability $1/4$, and doing k independent rounds drives the chance of misjudgment down to $\leq (1/4)^k$. Solovay and Strassen had earlier given a version with error $\leq (1/2)^{k20}$ (in 1977); Rabin's version is from 1980[19]. "Prime with probability $1 - \varepsilon$" is an epistemic object fundamentally different from "proven prime," but it is good enough for engineering, and it can be made as sharp as you like: just add a few more rounds.

What matters is to see clearly what has been given up here. What is given up is the kind of certainty, not the rigor. That bound $(1/4)^k$ is itself a theorem, proved tight. You have not lowered the standard; you have only swapped one standard for another that can be delivered within budget. The probabilistic method has its standing in pure mathematics as well: Erdős's probabilistic method (Alon and Spencer wrote it into a classic[26]) can prove that some object exists by proving that the probability of its appearing

is positive, yet without handing you the object itself. Existence is proved, the construction absent.

This stance runs all the way to the belief in RH itself. Montgomery in 1973, studying the pair correlation of the zeros[30], derived and conjectured that the normalized pair-correlation function of the zeros takes the form

$$R_2(u) = 1 - \left(\frac{\sin \pi u}{\pi u}\right)^2 .$$

At Princeton, Dyson recognized at a glance that this is exactly the pair correlation of the eigenvalues of the random-matrix Gaussian unitary ensemble (GUE). Odlyzko in 1987 did the numerics with a vast number of zeros[33], verifying the agreement to a degree that takes the breath away. None of this is a proof, yet it is extremely strong evidence, leading mathematicians to believe RH, and to believe it in a way no different in essence from the way a physicist believes a law that has not yet been falsified. The interior of mathematics has, it turns out, grown its own method for forming belief amid the unverifiable.

How Mathematicians Judge

So we arrive at the genuinely human part of this chapter: when the oracle never comes, what does a person of judgment rely on to hold a belief and to decide where to push.

Mathematics is deductive on the outside and plausible on the inside. Pólya wrote *Mathematics and Plausible Reasoning*[2] precisely to describe how mathematicians, lacking a proof, weigh the import of a proposition through analogy, induction, and special cases; Hadamard surveyed the psychology of mathematical invention[3], recording the rhythm of incubation and sudden insight; Poincaré left us that famous moment[5], the instant his foot touched the step

of the omnibus, when the connection between Fuchsian functions and non-Euclidean geometry surged up without warning. These are not substitutes for proof, but the things a person actually relies on along that instrumentless stretch of road before the proof arrives.

Why do mathematicians believe RH before a proof appears? Because the evidence accumulates in every direction and corroborates itself: a vast number of zeros verified on the line, many equivalent forms none of which has collapsed, an analogue already conquered (the Weil conjectures, that is, the Riemann hypothesis over function fields, proved by Deligne), statistical behavior matching the predictions of random matrices exactly. No single one of these is a proof, yet together they constitute a disciplined belief.

The mathematical community has reflected on the standing of this belief too. Thurston's 1994 essay "On Proof and Progress in Mathematics"[21] argues that what mathematics advances is human understanding, not merely the inventory of formal proofs; the 1993 debate between Jaffe and Quinn over "theoretical mathematics"[22] was precisely asking to what extent conjecture-driven, evidence-first work counts as mathematics. Tao asked what good mathematics is[29], and not one item in his answer is "already proved." Set these side by side and the very ability to judge "this proposition is most likely true, and worth investing in" is itself a calibrated belief, which is exactly what Part IV will name head-on. A capable actor is not paralyzed when the oracle is absent, nor does he pretend to certainty; he holds a belief with its scale marked, and then acts all the same.

Where This Chapter Leads

Having hit the purest form of the unverifiable, the mathematician did not wait for a decision procedure. What he got was this: cer-

tificates (verify a slice, prove a bound), proxy substitution (swap the target for an equivalent statement, and own up to the fact that it is often faithful but no easier), probabilistic acceptance (give up the two-valued verdict, hold a calibrated probability and act), and the judgment to hold a belief before the proof.

None of these is an emergency measure peculiar to mathematics. Caging untrusted code in a sandbox, auditing an organization, inferring behind an interface the preferences a user has left unspoken: the things one reaches for are these same few, only in different jargon. Part III will lift each move out of the field where it grew, name it separately, and juxtapose it across domains; that comparison table is the payload of the whole book.

One plain word remains here. This book's own central claim, that "responses converge on the same small set," I cannot at this moment verify either. My belief in it is the same kind of thing as a mathematician's belief in RH: a belief built on cross-domain evidence, with its scale marked, yet without a proof. Chapter 14 will return to this, and let the book do, in person, the very thing it has been describing all along.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

1. G. Polya (1945). *How to Solve It: A New Aspect of Mathematical Method*. Princeton University Press. [1] Pólya breaks mathematical problem-solving into four stages: understanding the problem, devising a plan, carrying it out, and looking back, and systematically lays out heuristic

strategies such as analogy, special cases, working backward, and auxiliary problems. What he writes is not the proof of theorems but the instrumentless process of exploration before a proof is found, exactly what the section "How Mathematicians Judge" cares about.

2. G. Polya (1954). *Mathematics and Plausible Reasoning* (2 vols.). Princeton University Press. [1][4] The two volumes treat, respectively, induction and analogy in mathematics and the logical structure of plausible reasoning, arguing that before they reach a rigorous proof mathematicians form their belief in a proposition by observing special cases, inducing patterns, and weighing evidence. The main text borrows from it to point out that "mathematics is deductive on the outside and plausible on the inside," and it is the source reading for understanding the concept of plausible reasoning.
3. J. Hadamard (1945). *An Essay on the Psychology of Invention in the Mathematical Field*. Princeton University Press. [1] Hadamard surveyed the creative psychology of mathematicians and distilled the rhythm of discovery into preparation, incubation, illumination, and verification, stressing subconscious work and the unheralded flash of inspiration. It provides a first-hand psychological study for this chapter's account of Poincaré-style insight, showing that much of mathematical judgment happens outside consciousness and outside proof.
4. H. Poincaré (1902). *La Science et l'Hypothèse*. Flammarion. [1] In this classic of the philosophy of science, Poincaré discusses the standing of mathematical hypotheses, conventions, and geometry, arguing that many foundational choices are not imposed by experience but adopted out of convention and convenience. It shows how a leading mathematician reflects on the epistemological foundations of his own discipline, in tune with this chapter's concern

with how to hold a belief where verification is unavailable.

5. H. Poincaré (1908). *Science et Méthode*. Flammarion. [1] The passage in which inspiration arrives as he steps onto the running board of an omnibus is the most often cited first-hand record in the psychology of mathematical discovery, by which Poincaré dissects the role of intuition, choice, and the subconscious in creation. The main text uses this moment directly, to show how insight connects scattered threads without any prior sign.
6. G. H. Hardy (1940). *A Mathematician's Apology*. Cambridge University Press. [1] Hardy defends the value of pure mathematics, holding that good mathematics lies in its seriousness, depth, and inevitable beauty, not in utility. As the introspection of a great number theorist on his own craft, it defines on what grounds a mathematician judges whether a piece of work is worth doing, of a piece with the question that closes this chapter, "what is good mathematics."
7. E. P. Wigner (1960). "The Unreasonable Effectiveness of Mathematics in the Natural Sciences." *Communications on Pure and Applied Mathematics*, 13(1). [2][3] Wigner marvels that abstract mathematical concepts can describe the physical world with such precision, calling this fit a strange gift we neither understand nor deserve. The puzzle this short essay poses has no agreed answer to this day; for this chapter it exemplifies how a belief in a deep regularity can be taken seriously in the absence of proof.
8. I. Lakatos (1976). *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge University Press. [1][3] Lakatos, taking Euler's polyhedron formula as his example, re-enacts through a fictional classroom dialogue how definition, proof, and counterexample mutually revise one another and together advance mathematics. It overturns the stereotype of mathematics as once-and-for-all deduction, presenting knowledge as growing crookedly through conjecture and

refutation, exactly suited to this chapter's concern with the true shape of mathematical progress.

9. B. Riemann (1859). "Über die Anzahl der Primzahlen unter einer gegebenen Größe." *Monatsberichte der Berliner Akademie*. [2][3] Riemann's eight-page paper continues the ζ function onto the complex plane, gives the functional equation, and links the distribution of primes to the nontrivial zeros of ζ; the conjecture about the location of the zeros written down there in passing is the later Riemann hypothesis. It is the source text of the whole chapter, and the line quoted at this chapter's opening about "set aside provisionally after futile attempts" comes from here.
10. H. M. Edwards (1974). *Riemann's Zeta Function*. Academic Press. [2] Edwards's monograph develops around Riemann's original 1859 paper, gradually laying out the theory of the ζ function, the prime number theorem, and the origins of the Riemann hypothesis, attending to both historical context and technical detail. It is the classic introduction to the ζ function and RH, providing reliable background for the concepts this chapter touches, such as zeros and the critical line.
11. E. C. Titchmarsh, rev. D. R. Heath-Brown (1986). *The Theory of the Riemann Zeta-function* (2nd ed.). Oxford University Press. [2] This is the standard advanced monograph on the analytic theory of the ζ function, systematically treating the distribution of zeros, zero-density estimates, and mean-value theorems on the critical line, with Heath-Brown's revision adding more recent advances. It represents the sum of the technical results rigorously established around RH, and is the specialist background reference when this chapter discusses "certificates and bounds."
12. E. Bombieri (2000). "Problems of the Millennium: The Riemann Hypothesis." Clay Mathematics Institute. [2][4] This is the official problem statement of RH written for the

Clay Mathematics Institute's Millennium Prize Problems, in which Bombieri concisely sets out the conjecture's origins, precise formulation, and weight in number theory. It is the authoritative entry point for understanding why RH is ranked among the problems of the century, and this chapter's assessment of RH's standing finds its warrant here.

13. J. B. Conrey (2003). "The Riemann Hypothesis." *Notices of the American Mathematical Society*, 50(3). [2][3] Conrey's survey, aimed at a broad readership, marshals the various kinds of evidence supporting RH, including the numerical verification of zeros, the agreement with random matrix theory, and the proven analogue over function fields. It gathers in one place the evidence on which this chapter's three stances rely, and is a convenient read for understanding why the mathematical community believes RH.
14. X.-J. Li (1997). "The Positivity of a Sequence of Numbers and the Riemann Hypothesis." *Journal of Number Theory*, 65(2). [2][4] Xian-Jin Li proves that RH is equivalent to a sequence of real numbers λ_n defined from the zeros being nonnegative for all n, translating the geometric statement about the location of the zeros into a positivity criterion for a sequence. The main text takes it as the foremost example of proxy substitution, showing how an equivalent rewriting can be faithful yet not necessarily easier.
15. E. Bombieri & J. C. Lagarias (1999). "Complements to Li's Criterion for the Riemann Hypothesis." *Journal of Number Theory*, 77(2). [2][4] The two authors point out that Li's criterion is in fact a special case of a family of general inequalities for an arbitrary multiset of complex numbers, not specific to the ζ function, and, using the Guinand-Weil explicit formula, give an arithmetic expression for λ_n, connecting it to Weil's criterion for RH. It deepens the understanding of Li's criterion and shows how the same equivalent proposition is rewritten again and again across different languages,

an extension of this chapter's theme of proxy substitution.

16. L. Báez-Duarte (2003). "A Strengthening of the Nyman-Beurling Criterion for the Riemann Hypothesis." *Atti della Accademia Nazionale dei Lincei, Rendiconti Lincei Mat. Appl.*, 14(1). [2][4] Báez-Duarte tightens the Nyman-Beurling approximation criterion into a sequence version using only integer dilations, making RH equivalent to a sequence of approximation distances d_n tending to zero. It is yet another equivalent rewriting listed in this chapter, carrying the zero problem into the framework of L^2 approximation, and likewise confirms that a faithful proxy is often no easier to solve.
17. X. Gourdon (2004). "The 10^13 First Zeros of the Riemann Zeta Function, and Zeros Computation at Very Large Height." Online technical report (numbers.computation.free.fr). [2][4] Gourdon, using high-precision floating-point computation with the Odlyzko-Schönhage algorithm, verified that the first 10^{13} zeros all fall on the critical line. This chapter deliberately sets it against Platt: it gives extremely strong numerical confidence but does not rigorously bound the rounding error, and so it is a numerical result rather than a mechanically recheckable certificate.
18. D. J. Platt (2017). "Isolating Some Non-trivial Zeros of Zeta." *Mathematics of Computation*, 86(307). [2][4] Platt uses interval arithmetic to isolate the zeros rigorously on the critical line, giving the error a provable upper bound and thereby raising numerical verification to a mechanically recheckable certificate. This chapter uses it to exemplify the stance of "certificates and bounds": prove not the whole, but a guaranteed bound for a finite slice.
19. M. O. Rabin (1980). "Probabilistic Algorithm for Testing Primality." *Journal of Number Theory*, 12(1). [2][4] Rabin gives the Miller-Rabin primality test: if n is composite, a

randomly chosen base slips past it with probability at most $1/4$, and k independent rounds drive the chance of misjudgment down to $(1/4)^k$. This chapter uses it as the cleanest example of the probabilistic method, showing that the error bound itself is a rigorously proved theorem, and what is given up is the kind of certainty, not the rigor.

20. R. Solovay & V. Strassen (1977). "A Fast Monte-Carlo Test for Primality." *SIAM Journal on Computing*, 6(1). [2][4] Solovay and Strassen earlier proposed a probabilistic primality test based on the Jacobi symbol, with a single-round chance of misjudgment of at most $1/2$, one of the founding works of randomized algorithms. This chapter places it alongside Rabin's version, showing that trading the probabilistic method for a decision deliverable within budget has long had precedents in computational number theory.
21. W. P. Thurston (1994). "On Proof and Progress in Mathematics." *Bulletin of the American Mathematical Society*, 30(2). [1][3] Thurston argues that what mathematics truly advances is human understanding of mathematics, not merely the inventory of formal proofs, and that proof is only a socialized means by which the community transmits and confirms understanding. This chapter cites it at the close to challenge the narrowing view that "mathematics equals proven theorems," and it is essential reading for reflecting on the standing of proof.
22. A. Jaffe & F. Quinn (1993). " "Theoretical Mathematics": Toward a Cultural Synthesis of Mathematics and Theoretical Physics." *Bulletin of the American Mathematical Society*, 29(1). [1][3] Jaffe and Quinn propose distinguishing "theoretical mathematics" from rigorous mathematics, suggesting that conjecture-driven, not-yet-rigorously-proved work be explicitly labeled so as not to erode the reliability of mathematics, which set off a widely noted debate in the community. This chapter borrows from that debate to ask

to what extent evidence-first work counts as mathematics, cutting right to the whole book's concern with verification and belief.

23. J. von Neumann (1947). "The Mathematician." In R. B. Heywood (ed.), *The Works of the Mind*. University of Chicago Press. [1][3] In this essay von Neumann reflects on the nature of mathematics, on how mathematics travels back and forth between abstraction and empirical sources, how it chooses its direction by aesthetic standards, and why drifting too far from the empirical source carries the risk of degeneration. From the vantage of a master who ranged across many fields, it shows the weight of aesthetics and taste in mathematical judgment, echoing this chapter's discussion of how mathematicians decide where to push.
24. K. Appel & W. Haken (1977). "Every Planar Map Is Four Colorable, Part I: Discharging." *Illinois Journal of Mathematics*, 21(3). [2][3] Appel and Haken, with the aid of a computer-checked set of unavoidable configurations, proved the four color theorem, the first famous mathematical proof to depend essentially on a computer. It raised a debate that continues to this day: whether a proof no human can read line by line still counts as a proof, directly bearing on this chapter's discussion of certificates and mechanically recheckable guarantees.
25. T. Hales et al. (2017). "A Formal Proof of the Kepler Conjecture." *Forum of Mathematics, Pi*, 5. [2][3][4] Hales's team, in the Flyspeck project, used the HOL Light and Isabelle proof assistants to complete a fully formalized, mechanically checkable proof of the Kepler conjecture, settling the unresolved status the original proof had been left in because human referees could not fully check it. This chapter uses it to show how formal verification compresses a contested argument into a line-by-line verifiable certificate.
26. N. Alon & J. H. Spencer (1992). *The Probabilistic Method*.

Wiley. [2][4] This classic systematically presents the probabilistic method pioneered by Erdős: to prove that some combinatorial object exists, prove that the probability of its appearing at random is positive, and thereby conclude that it must exist, while often being unable to construct it explicitly. This chapter borrows from it to point out the feature of the probabilistic method in proving existence in pure mathematics: existence proved, construction absent.

27. P. J. Davis & R. Hersh (1981). *The Mathematical Experience*. Birkhäuser. [1][3] Davis and Hersh, starting from the actual experience of mathematicians, discuss the existential status of mathematical objects, the role of proof, and the philosophical situation of mathematics, presenting a practitioner's perspective different from formalist dogma. It provides this chapter with a close-to-the-ground reflection for understanding how mathematicians hold beliefs in practice and regard truth.
28. W. T. Gowers (2000). "The Two Cultures of Mathematics." In *Mathematics: Frontiers and Perspectives*. American Mathematical Society. [1][3] Gowers distinguishes two cultures within mathematics: theory-builders and problem-solvers, the former solving problems for the sake of understanding, the latter understanding for the sake of solving problems, with algebraic geometry, the Langlands program, and combinatorial number theory as contrasts. It shows that mathematicians can hold different measures of what is deep and what is good work, echoing this chapter's discussion of the standards of mathematical judgment.
29. T. Tao (2007). "What Is Good Mathematics?" *Bulletin of the American Mathematical Society*, 44(4). [1][3] Tao enumerates the many mutually distinct dimensions of good mathematics, from rigor, depth, and beauty to richness of application and the power to open new directions, arguing that no single standard exists and that, over the long run,

these qualities often pull on one another. This chapter borrows from it to show that judging whether a piece of work is worth investing in is itself an ability, and that not one of its standards is "already proved."

30. H. L. Montgomery (1973). "The Pair Correlation of Zeros of the Zeta Function." In *Analytic Number Theory*, Proc. Sympos. Pure Math., XXIV. American Mathematical Society. [2][3] Montgomery studies the normalized pair correlation of the zeros of ζ, deriving and conjecturing its form, and Dyson at once recognized this as exactly the pair correlation of the eigenvalues of the random-matrix Gaussian unitary ensemble. This linkage opened the deep connection between number theory and random matrix theory, and is one of the key pieces of evidence when this chapter discusses on what the belief in RH is built.
31. P. Sarnak (2004). "Problems of the Millennium: The Riemann Hypothesis." Clay Mathematics Institute. [2][3][4] Sarnak's note for the Clay Institute emphasizes the generalized forms of RH and its central role in analytic number theory, explaining why so many other results take it as a premise. From the vantage of an expert active in the connection between zero statistics and random matrices, it rounds out this chapter's understanding of RH's importance and its web of evidence.
32. J. C. Lagarias (2002). "An Elementary Problem Equivalent to the Riemann Hypothesis." *The American Mathematical Monthly*, 109(6). [2][4] Lagarias gives an elementary inequality using only the harmonic numbers and the sum-of-divisors function, proving that it holds for all n if and only if RH holds, translating the abstruse zero problem into an arithmetic statement that could almost be written on a postcard. This chapter uses it to show that proxy substitution can be elementary on the surface while the difficulty remains undiminished.

33. A. M. Odlyzko (1987). "On the Distribution of Spacings Between Zeros of the Zeta Function." *Mathematics of Computation*, 48(177). [2][4] Odlyzko, with a vast amount of high-precision computation, examines the distribution of spacings between the zeros of ζ and finds it in astonishing agreement with the prediction of the random-matrix Gaussian unitary ensemble, providing strong numerical support for the Montgomery-Dyson conjecture. This chapter uses it to show that this statistical fit, though not a proof, is extremely strong evidence leading mathematicians to believe RH.

Chapter 8: The Organization Blind to Itself

Thesis: A large organization or a state cannot directly observe its own knowledge and activity, which is distributed, partly hidden, and at times strategically concealed; so it reaches out to grasp a proxy (the metric it can see), and this is where the proxy move's Goodhart failure is at its most glaring, before being shored up with auditing (the audit trail) and redundancy.

A Colossus That Cannot See Itself

The principal-agent problem of the previous chapter now has its scale enlarged. The principal is no longer a single person, but an entire organization, a whole state; the ones entrusted to act are thousands upon thousands of people scattered everywhere. A new, almost absurd situation arises: this colossus cannot see itself clearly.

It wants to know how many people it has, what is being planted, who is doing what, and how well they are doing it; yet none of

this knowledge resides anywhere it can directly read. What this chapter examines is this: when the object to be verified is the organization's own knowledge, distributed, hidden, and apt to dodge being seen, what does the organization do? The unverifiability here lays several of the earlier faces one atop another: partial observability (knowledge dispersed at the edges), plus the adversarial (the people being watched will turn around and manipulate the thing being watched).

Distributed Knowledge

Hayek's 1945 essay "The Use of Knowledge in Society"[1] laid the problem bare: the knowledge on which a society runs is never concentrated in any one place; it is dispersed among countless individuals, it is local knowledge concerning a particular time and a particular place, and often it cannot even be put into words. Which machine has a small fault today, which customer is in fact about to drift away, which side path will collapse after the rain, the holders of such knowledge often do not themselves realize it is "knowledge," let alone find a way to package it up and hand it to the center. Polanyi called this layer the tacit dimension[2]: we know far more than we can tell.

This means the organization faces more than a case of "the information has not yet been collected." Even if everyone were loyal and cooperative, that local, tacit knowledge would still evaporate in the course of being gathered. The "whole picture of the organization" the center wants cannot, in principle, be faithfully fitted into any container that could verify it. This is the social-scale version of partial observability, and it comes with a harder floor: such knowledge is by its very nature local, and cannot be gathered into any one place.

The Urge Toward Legibility

If it cannot be seen clearly, the impulse is to find a way to make it seeable. Scott's 1998 book *Seeing Like a State*[3] gives this urge an exact name: legibility. For a state to act upon society, it must first remake society into a shape it can read. It measures the land and draws cadastral maps, imposes fixed surnames on people who once had only nicknames, bynames, or patronymics, standardizes weights and measures, and rolls out standardized scientific forestry. These are not neutral acts of recording; they are reshaping reality itself so that reality will fit into the table. Hacking's *The Taming of Chance*[32], Desrosières's *The Politics of Large Numbers*[31], and Bowker and Star's *Sorting Things Out*[30], taken together, form a history of "making society countable."

The danger of legibility lies in the fact that the map must simplify, and once an organization acts only according to the map, the things the map has erased come back to bite. Scott's most forceful case is precisely scientific forestry: to render the forest "legible, countable, and taxable," the Prussians remade the tangled natural woodland into uniform, easily inventoried single-species plantations. The first generation or two grew splendidly; by the third generation the soil was exhausted, pests spread, and the forest died off in swaths, so much so that German even coined a word for it, Waldsterben, the death of the forest. The cleaner the map, the more lethal the local knowledge it erased, the knowledge that had kept the system running. This is the organization manufacturing for itself the observability it lacks, at the cost of leveling, with its own hands, the very complexity that let it run.

The Proxy Metric, and Its Goodhart Collapse

The most common landing place of the urge toward legibility is the metric. The things truly cared about, health, learning, productivity, public welfare, cannot be directly observed; so the organization grasps the proxy it can see, the KPI, GDP, exam scores, paper citation counts, emergency-room waiting times.

This is exactly the proxy substitution we met in Chapter 7. But here it fails in the opposite manner, and that contrast is one of this book's main threads. The mathematician's proxy is faithful but no easier: an equivalent rewriting really is equivalent, yet it has not become any simpler to solve. The organization's proxy is precisely the reverse, easier but unfaithful: the metric is of course easy to measure, but its correspondence to the true target snaps the moment the metric itself becomes the target.

This rupture goes by many names. Goodhart in 1975[4]: once a metric is taken as a policy target, its reliability as a metric falls apart. Campbell in 1979[6] says the social version of the same thing. As early as 1956, Ridgway had catalogued the "dysfunctional consequences of performance measurements"[7]; Kerr's 1975 essay "On the Folly of Rewarding A, While Hoping for B"[8] wrote it into the common sense of management. Strathern gave it its most distilled single sentence[5]: when a measure becomes a target, it ceases to be a good measure. A deeper layer is reflexivity: Espeland and Sauder in 2007[12] point out that a public ranking is not describing the world but remaking it; a ranked university will change itself to fit the ranking's algorithm, so that what the metric "measures" is precisely the behavior it has itself called into being. Bevan and Hood[11] documented the gaming of metrics inside the English health system; Smith in 1995[10] analyzed how the public release of performance data invites a string of unforeseen consequences; and Merton's 1936 paper on "the unanticipated consequences of pur-

posive social action"[9] is the wellspring of all this. Such collapses are common in reality, and the cost is at times staggering. To hit its account metric for "cross-selling," Wells Fargo employees secretly opened some 3.5 million fake accounts without customers' knowledge; the affair came to light in 2016, the bank was fined 185 million dollars and more than five thousand employees were dismissed, and the number that had been enshrined had destroyed the very customer relationship it was meant to measure. An earlier, parable-like case took place in colonial-era Delhi: to wipe out snakes, the authorities offered a bounty for dead cobras, whereupon residents simply bred cobras to collect it; when the bounty stopped, the snakes were all set free, and the snake problem grew worse than before. The "cobra effect" takes its name from this.

Why must a proxy be distorted? Principal-agent theory gives the rigorous explanation. Holmström's 1979 informativeness principle[14]: reward should be hung on signals that carry information about "effort." But once effort is multidimensional and you can measure only a few of its dimensions, trouble begins. Holmström and Milgrom's 1991 multitask analysis[15] (the multitask principal-agent model) puts it plainly: when a person must attend to measurable and unmeasurable tasks at once, the more heavily you reward the measurable part, the more they will shift effort away from the unmeasurable part toward the measurable. Let the true target be G and the observable proxy be P; the two are correlated under the status quo. The problem is that this correlation is a product of behavior, not an objective law. Once P is made the target of pressure,

$$\arg\max_a P(a) \qquad \text{vs.} \qquad \arg\max_a G(a),$$

the rational agent goes looking for actions that raise P while doing nothing for, or even harming, G; the correlation is crushed by the very pressure to optimize. The teacher teaching to the test, the

hospital scheduling patients so as to lower one particular waiting-time figure, the researcher slicing a single paper into the smallest countable units of publication, are all the same mechanism.

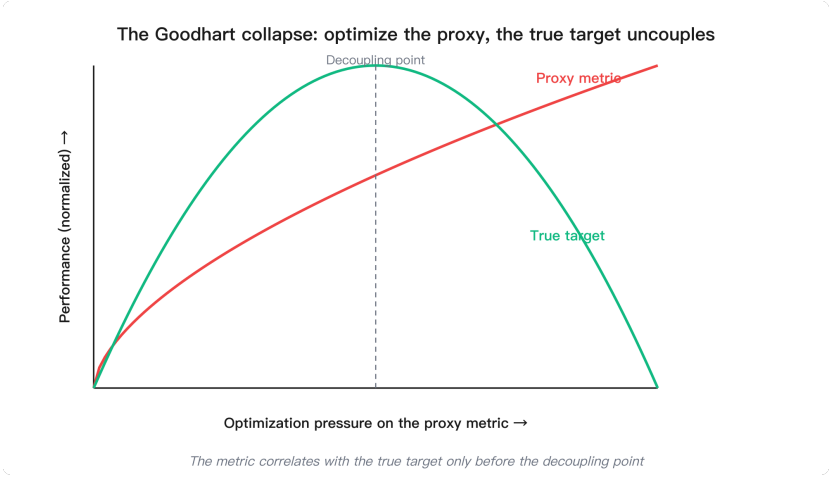

Figure 6: The Goodhart collapse: optimize the proxy, and the true target comes uncoupled

Shoring It Up With Auditing and Redundancy

The proxy on its own will collapse, so the organization adds two further moves, which are likewise moves that recur throughout this book.

The audit trail and auditing. Double-entry bookkeeping is one of humanity's oldest audit chains; Soll, in *The Reckoning*[22], argues that the ability to keep accounts that can be checked bears directly on the rise and fall of one nation after another: those that can reckon themselves are the ones that endure. Modern financial auditing and independent inspection all amount to swapping "fraud cannot be prevented in advance" for "fraud can be detected after the fact." But this move has an ailment of its own. Power's

1997 *The Audit Society*[20] spells it out: when verification itself becomes a ritual, what the organization produces is no more than the appearance that "everything is under control," rather than control itself. The "audit culture" of Shore and Wright[17] and O'Neill's reflection on "trust" in the 2002 Reith Lectures[21] speak of the same alienation: in order to be held accountable, institutions pour enormous effort into manufacturing traces that can be inspected, while the real work gets pushed to the side.

Redundancy and consensus. Landau's underrated 1969 article[16] rehabilitated "duplication and overlap": in a system whose parts are none of them fully reliable, redundancy is not waste but a source of reliability; several mutually independent checks are harder to fool all at once than a single authority. This move holds only under one precondition, independence, which the next part will stress again and again: if several checks are in fact of the same origin, a correlated failure will destroy the entire value of the redundancy in one stroke.

Where This Chapter Leads, and the Close of Part II

With this, the four sites have all been surveyed. The person at the console, the agent set loose, the mathematician hitting the wall, and the organization blind to itself face sources of unverifiability that are wildly different: preferences hidden in the heart, future behavior in an open world, propositions undecidable in principle, knowledge that is distributed and apt to dodge. Yet what they reach out to grasp is the same small set of things.

What most deserves to be set side by side is the two opposite failures of proxy substitution. The mathematician comes to grief on faithful but no easier; the organization comes to grief on easier but unfaithful. The two ends of that 2×2 table in Chapter 7

now both have flesh on them. They are not two moves but two directions of failure of a single move, and a good proxy must dodge both ends at once, being faithful and easier alike, which is so rare that it is nearly the whole of the craft. Chapter 11 will formally join these two ends. The principal-agent skeleton, too, has grown from a snippet of code in Chapter 6 into a state here.

With this, Part II has demonstrated the moves "embedded in their sites and tangled together." They are scattered, go by changing names, and are mixed into their respective jargons. What Part III sets out to do is to pull each move out of the field in which it grew, wash it clean, name it on its own, and cover every site at one stroke. That comparative table is the true payload of this book.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

1. F. A. Hayek (1945). "The Use of Knowledge in Society." *American Economic Review*, 35(4), 519-530. [2][4] Hayek argues that the knowledge on which a society runs is never concentrated in one place but dispersed among countless individuals, that it is local knowledge concerning a particular time and a particular place, and that it cannot be faithfully gathered by any center. This essay is the direct point of departure for this chapter's section "Distributed Knowledge," and it sets the epistemological coloring of the predicament that "the organization cannot see itself clearly."
2. M. Polanyi (1966). *The Tacit Dimension*. Doubleday. [2]

Polanyi proposes the tacit dimension of knowledge, his famous line being "we know more than we can tell." The book uses it to show that a considerable part of the local knowledge dispersed at the edges simply cannot be put into words and handed up, which is a harder floor beneath the organization's difficulty in verifying itself.

3. J. C. Scott (1998). *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press. [2][4] Scott proposes the concept of "legibility": in order to act upon society, a state uses such means as cadastral maps, fixed surnames, and standardized weights and measures to remake society into a shape it can read, and this simplification often erases the local knowledge that keeps the system running, leading to failures like scientific forestry. This chapter's section "The Urge Toward Legibility" is built precisely on this; it is the core reading for understanding why an organization sets about leveling complexity with its own hands.
4. C. A. E. Goodhart (1975). "Problems of Monetary Management: The U.K. Experience." *Papers in Monetary Economics*, Vol. I. Reserve Bank of Australia. [2] Goodhart was originally speaking of monetary policy, yet he gave an insight later cited everywhere: once a statistical regularity is taken as the target of policy control, its original regularity falls apart. This is the source of the name of this chapter's section "The Proxy Metric, and Its Goodhart Collapse," and the starting point for understanding how a proxy is crushed by the pressure to optimize.
5. M. Strathern (1997). "Improving Ratings: Audit in the British University System." *European Review*, 5(3), 305-321. [2][4] Strathern, drawing on the experience of auditing in British universities, left Goodhart's law its most distilled popular formulation: when a measure becomes a target, it ceases to be a good measure. This chapter quotes the sen-

tence directly; it is also the single best line for explaining the abstract collapse of the proxy to a reader.

6. D. T. Campbell (1979). "Assessing the Impact of Planned Social Change." *Evaluation and Program Planning*, 2(1), 67-90. [2][4] Campbell, from the standpoint of social-science evaluation, proposed "Campbell's law," isomorphic to Goodhart's: the more a quantitative social indicator is used for social decision-making, the more it is subject to corruption pressures, and the more it will distort the very social process it was meant to monitor. This chapter uses it to corroborate that the collapse of the proxy is not peculiar to economics but the same phenomenon discovered again and again across disciplines.
7. V. F. Ridgway (1956). "Dysfunctional Consequences of Performance Measurements." *Administrative Science Quarterly*, 1(2), 240-247. [2][4] Ridgway, very early on, systematically catalogued the dysfunctional consequences of performance measurement, distinguishing the distortions brought by single, composite, and multiple measures. This chapter uses it to show that the gaming and distortion of metrics is a rather old problem, discovered quite early, and not some recent coinage of management studies.
8. S. Kerr (1975). "On the Folly of Rewarding A, While Hoping for B." *Academy of Management Journal*, 18(4), 769-783. [2][4] Kerr lists a wealth of real-world examples to show that organizations often reward one kind of behavior while hoping for another that they have not rewarded, with results that naturally run counter to intent. This essay wrote the mismatch of proxy and incentive into the common sense of management, and is the classic source of this chapter's "reward A while hoping for B" mechanism.
9. R. K. Merton (1936). "The Unanticipated Consequences of Purposive Social Action." *American Sociological Review*, 1(6), 894-904. [2][4] Merton systematically analyzed why

purposive social action always brings unanticipated consequences, and sorted out their causes, such as ignorance, error, and the imperious immediacy of value. This chapter treats it as the wellspring of a whole series of "unforeseen" phenomena, such as the gaming of metrics and the backlash of legibility.

10. P. Smith (1995). "On the Unintended Consequences of Publishing Performance Data in the Public Sector." *International Journal of Public Administration*, 18(2-3), 277-310. [2][4] Smith classified and sorted out the string of unintended consequences invited by the public release of performance data in the public sector, such as tunnel vision, myopia, misrepresentation, measure fixation, and gaming. This chapter uses it to break the vague "the metric gets distorted" into several recognizable, concrete modes of failure.
11. G. Bevan & C. Hood (2006). "What's Measured Is What Matters: Targets and Gaming in the English Public Health Care System." *Public Administration*, 84(3), 517-538. [2][4] Bevan and Hood empirically documented the various ways of gaming metrics in the English National Health Service under "targets and terror" governance, such as scheduling patients so as to lower waiting times, a practice that appeases the metric while doing nothing for real health. This chapter takes it as field evidence of how the gaming of metrics actually happens in public services.
12. W. N. Espeland & M. Sauder (2007). "Rankings and Reactivity: How Public Measures Recreate Social Worlds." *American Journal of Sociology*, 113(1), 1-40. [2][4] Espeland and Sauder, using law-school rankings as their example, propose "reflexivity": a public measure does not merely describe the world but turns around to reshape the behavior of those being measured, so that what the metric finally measures is the very reaction it has itself called into being. This chapter's paragraph "A deeper layer is reflexivity"

comes from here; it pushes the failure of the proxy to the level where the metric manufactures reality.

13. M. Sauder & W. N. Espeland (2009). "The Discipline of Rankings: Tight Coupling and Organizational Change." *American Sociological Review*, 74(1), 63-82. [2][4] This companion piece draws on Foucault's concept of discipline to analyze how rankings become embedded in organizations: institutions once loosely coupled are forced under the pressure of rankings into tight coupling, and the external measure is internalized as everyday self-surveillance and organizational change. It complements the previous entry, which treats the mechanism of reflexivity, while this one treats how rankings remake an organization's internal structure.
14. B. Holmström (1979). "Moral Hazard and Observability." *The Bell Journal of Economics*, 10(1), 74-91. [2][4] Holmström proposes the informativeness principle: under moral hazard, the optimal reward contract should hang on all signals that carry information about the agent's effort. This chapter uses it to give a rigorous principal-agent explanation of "why the proxy is bound to be distorted," and to lead into the trouble that arises when effort is multidimensional and only a few dimensions can be measured.
15. B. Holmström & P. Milgrom (1991). "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *The Journal of Law, Economics, and Organization*, 7(Special Issue), 24-52. [2] The multitask principal-agent model shows that when a person must attend to measurable and unmeasurable tasks at once, the more heavily the measurable part is rewarded, the more they will draw effort away from the unmeasurable part. This chapter argues from it the mechanism of the proxy's collapse: pressing on the observable metric rationally induces the agent to abandon work that is hard to measure yet truly important.

16. M. Landau (1969). "Redundancy, Rationality, and the Problem of Duplication and Overlap." *Public Administration Review*, 29(4), 346-358. [2][4] Landau rehabilitated the "duplication and overlap" so often denounced as waste: in a system whose parts are none of them fully reliable, redundancy is precisely the source of reliability, and several mutually independent checks are harder to fool all at once than a single authority. This chapter's section "Shoring It Up With Auditing and Redundancy" adopts this argument directly, and stresses that its precondition is the mutual independence of the checks.
17. C. Shore & S. Wright (1999). "Audit Culture and Anthropology: Neo-Liberalism in British Higher Education." *The Journal of the Royal Anthropological Institute*, 5(4), 557-575. [2][4] Shore and Wright, taking British higher education as their example, propose "audit culture": under neoliberal governance, the logic of accountability and audit seeps into academic institutions, turning peers into objects of surveillance and reshaping the way people govern themselves. This chapter uses it to show how auditing is alienated from a tool into a culture that makes people spend themselves on manufacturing inspectable traces.
18. J. Z. Muller (2018). *The Tyranny of Metrics*. Princeton University Press. [4] Muller, writing for the general reader, surveys the distortions and costs brought by overreliance on quantitative metrics in fields such as medicine, education, policing, and business, and offers judgments about when measurement should and should not be used. The book is a popular synthesis that explains the collapse of the proxy to practitioners, suitable for the reader as an introduction and a point of comparison.
19. T. M. Porter (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press. [2][4] Porter argues that reliance on quantification

often springs from a kind of "mechanical objectivity": in situations lacking trust and demanding outward accountability, numbers are used as a tool to suppress personal judgment and ward off challenge. The book provides a deep sociological explanation of why organizations cling to legible numbers, serving as background to both this chapter's sections on legibility and on auditing.

20. M. Power (1997). *The Audit Society: Rituals of Verification*. Oxford University Press. [2][4] Power points out that when verification itself becomes a set of rituals, what the organization produces is often the appearance that "everything is under control," rather than control itself, and society remakes itself in turn so as to be auditable. This chapter's section "Shoring It Up With Auditing and Redundancy" draws on it to spell out the ailment that the audit move carries within: the more traces, the more the real work gets pushed aside.
21. O. O'Neill (2002). *A Question of Trust: The BBC Reith Lectures 2002*. Cambridge University Press. [4] O'Neill, in this set of Reith Lectures, reflects on the contemporary culture of accountability: the various measures of transparency and audit meant to rebuild trust often erode the very trust they were intended to foster, leaving people busy coping with inspection rather than doing the work well. This chapter cites it alongside "the audit society" to show how excessive accountability backfires.
22. J. Soll (2014). *The Reckoning: Financial Accountability and the Rise and Fall of Nations*. Basic Books. [1][4] Soll, with double-entry bookkeeping as his thread, argues that the ability to keep accounts that can be checked bears directly on the rise and fall of one nation after another: those that can reckon themselves are the ones that endure. This chapter uses it to support the claim that "the audit trail and auditing" is one of humanity's oldest chains of verification.
23. J. G. March & H. A. Simon (1958). *Organizations*. John

Wiley & Sons. [2] March and Simon laid the foundations of modern organization theory: the rationality of an organization's members is bounded, and the organization copes with the limits of individual cognitive capacity precisely through division of labor, procedures, and information channels. The book provides a basic framework for "the organization cannot see itself," and is a classic source for understanding how information flows and decays through a hierarchy.

24. H. A. Simon (1947). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*. Macmillan. [2] Simon proposes bounded rationality, understanding the organization as a set of structures that help members make decisions under limited cognitive capacity. The book is the source for understanding why an organization must rely on simplification, routine, and proxies to run, and it lays the theoretical bedrock for this chapter's account of the limits of organizational self-knowledge.
25. R. M. Cyert & J. G. March (1963). *A Behavioral Theory of the Firm*. Prentice-Hall. [2] Cyert and March propose a behavioral theory of the firm, stressing that organizational decisions are governed by standard operating procedures, limited search, and the negotiation of goals among parties, rather than by pure optimization. The book helps in understanding the plurality and tension of goals within an organization, and is an important support for this chapter's treatment of the organization as a bounded-rationality actor.
26. O. E. Williamson (1975). *Markets and Hierarchies: Analysis and Antitrust Implications*. Free Press. [2] Williamson, starting from transaction costs, explains why some activities are coordinated by the market and others are folded into a hierarchical organization: bounded rationality and opportunism make some transactions more efficient to complete within a hierarchy. The book provides an economic explana-

tion of why an organization takes scattered activities under its own roof, and thereby shoulders the difficulty of verifying them.

27. K. J. Arrow (1974). *The Limits of Organization.* W. W. Norton. [2][4] Arrow concisely explores the organization as a means of coping with the scarcity of information and with uncertainty, and the inherent limits it meets in authority, responsibility, and trust. The book points out that trust is an indispensable lubricant of social functioning that cannot be bought by contract, in distant resonance with the cost of verification examined in this chapter's sections on auditing and redundancy.
28. M. Lipsky (1980). *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services.* Russell Sage Foundation. [2][4] Lipsky points out that street-level bureaucrats such as teachers, police, and social workers exercise a great deal of discretion under conditions of scarce resources, and that their everyday coping in fact shapes how public policy actually lands. The book is an important reference for understanding why the local knowledge and discretion at the organization's edges is hard for the center to observe and verify.
29. J. Q. Wilson (1989). *Bureaucracy: What Government Agencies Do and Why They Do It.* Basic Books. [2][4] Wilson examines in detail the actual workings of government agencies, distinguishing types of agency by whether outputs and outcomes are observable, and explains why the real effectiveness of many public agencies is hard to measure. The book provides rich real-world material for this chapter's "the organization blind to itself," and is especially helpful for understanding why proxy metrics are particularly apt to distort in the public sector.
30. G. C. Bowker & S. L. Star (1999). *Sorting Things Out: Classification and Its Consequences.* MIT Press. [2][4] Bowker

and Star examine how classification systems silently embed themselves into infrastructure, and how they shape the very reality they meant to record neutrally, with the differences flattened by classification often carrying real consequences. This chapter sets it alongside Hacking and Desrosières, gathering it into the history of "making society countable," to show that classification is an invisible link in the engineering of legibility.

31. A. Desrosières (1998). *The Politics of Large Numbers: A History of Statistical Reasoning* (trans. C. Naish). Harvard University Press. [2] Desrosières traces the history of statistical reasoning, showing that statistical categories took shape in step with state administration, and that numbers are at once a tool for knowing society and a political act that constructs social reality. This chapter brings it into the genealogy of "making society countable," revealing the provenance of the statistical apparatus behind legibility.
32. I. Hacking (1990). *The Taming of Chance*. Cambridge University Press. [2] Hacking examines the rise of statistical and probabilistic thought in the nineteenth century, arguing that the mass collection of population data "tamed chance" and gave birth to concepts such as "the normal" and "normalcy" that govern modern governance. This chapter cites it to show that making society countable is itself a stretch of history that remade cognition, and not a neutral act of recording.

Part III: Convergence

Chapter 9: Compressing the Unknown

> **Thesis:** Two moves attack uncertainty head-on. On a slice you can inspect, deliver a guaranteed bound (a certificate); spend your limited checks where they dissolve the most uncertainty (optimal screening).

Part II embedded the moves inside four concrete sites and let them play out, entangled with one another. From this part onward the angle changes: each move is lifted out on its own, washed clean of its domain's jargon, presented in pure form, and then laid across all the sites at once. This is the real payload of the book, that "table of the same move under its many vocabularies."

The eight moves pair off, two by two, into four chapters. The pairing is not a convenience; the pairing is itself a claim: each pair pulls on the same more basic lever, a point that comes due in Chapter 13. This chapter's pair, certificate and optimal screening, together attack the same thing, namely uncertainty. They compress the unknown from opposite ends: at one end, on a slice you can manage to check, you prove a guaranteed bound; at the other, you spend your limited checks where they dissolve the most uncertainty.

And, by the iron rule laid down in Chapter 4, every move proposed

and every cross-domain juxtaposition made must be interrogated once more: is this transfer substantive (same mechanism, same failure mode, same trade-off), or merely a pretty metaphor?

Certificate: Prove a Bound on a Slice

The first move in pure form: rather than verify the whole, produce a bounded, independently recheckable local guarantee. What you deliver is not "all of it is right" but "within this range, it is wrong by at most this much," accompanied by a voucher anyone can quickly verify.

Its cross-domain forms are astonishingly consistent.

In machine learning it is called a generalization bound. Valiant's 1984 PAC framework[1] and the VC dimension of Vapnik and Chervonenkis (brought in by Blumer and colleagues in 1989[4]) give a guarantee of the form "with probability at least $1-\delta$, the true error does not exceed the empirical error plus a complexity penalty":

$$R(h) \;\le\; \hat{R}(h) + \sqrt{\frac{d(\,\ln(2n/d)+1) + \ln(4/\delta)}{n}}.$$

You cannot verify the model's behavior on all future data (that is the open world), but on the slice of "samples already seen" you can prove a bound, with a confidence level, that holds for unseen data. Hoeffding's inequality[17] is its probabilistic engine; PAC-Bayes (McAllester[6]) is its refinement.

In software it is called types and proofs. A type system does not prove a program "all correct," only that one property holds (for instance, that it will not treat an integer as a pointer), and in exchange it buys a decidable, mechanically recheckable check (Pierce[7]). The Curry-Howard correspondence (Howard 1980[8]) sets an equals sign between "proof" and "program," and Necula's

1997 proof-carrying code[9] pushes the move to its limit: untrusted code carries its own safety proof, and the host need only quickly check the certificate, without re-deriving anything itself. Leroy's formally verified CompCert compiler of 2009[10] and the Z3 solver of de Moura and Bjørner in 2008[11] are the industrialization of the same idea.

In numerical computation it is called an error bound. Higham's 2002 backward error analysis[12] and Moore's 1966 interval arithmetic[13] let you compute carrying "an interval guaranteed to contain the true value," so that what you finally deliver is not a floating-point number that may deceive you, but a guaranteed range. In mathematics it is the zeros verified up to height T from Chapter 7: a bound, not a theorem.

The unifying idea is this: a certificate is a local, bounded, independently recheckable guarantee. Its finest feature is that it exploits the verification asymmetry from Chapter 2: producing a certificate may be extremely expensive, while checking one is extremely cheap. Proof-carrying code, the solution to an NP problem, a mathematical proof, all feed on this same dividend.

It has only one standard failure mode, but a common one: the vacuous bound. A guarantee that is true yet useless, such as "the error is at most one hundred percent" or "this model's generalization error is finite," is logically unimpeachable and operationally worthless. The value of a bound lies not in holding, but in being tight enough to act on.

Optimal Screening: Spend Your Checks on the Cutting Edge

The second move in pure form: information has a cost, so allocate your limited checks to where, at the margin, they compress the most uncertainty.

Its cross-domain forms are equally tidy. In statistics and science it is called design of experiments: Fisher's 1935 *The Design of Experiments*[19] and Box and colleagues' *Statistics for Experimenters*[20] teach how to wring the most information from the fewest trials; Lindley in 1956 gave a measure of "the information an experiment provides"[21], and Chaloner and Verdinelli systematized Bayesian experimental design[22]; Wald's 1945 sequential test[23] lets you decide whether to continue as the data come in. Shannon's 1948 information theory[14] is the underlying currency of all of it. In machine learning it is called active learning: which sample is most worth labeling next (Cohn and colleagues[33], Settles[34]). In software testing it is called fuzzing: which input to throw compute at to knock out a crash (Miller and colleagues' pioneering 1990 experiment[35]). The modern scale of this move is striking. Since 2016 Google's OSS-Fuzz has continuously fed vast quantities of malformed input into thousands of open-source projects, and has by now uncovered tens of thousands of defects and vulnerabilities. No human testing team could exhaust to that magnitude; it relies precisely on pouring compute, without pause, toward the places most likely to crash. In auditing it is called sampling: which transactions to check to most likely find a problem. In interfaces it is which question to ask the user (Chapter 5).

Behind all of these lies the same optimization problem: maximize the expected information gain between the answer and the unknown,

$$q^\star = \arg\max_q \; I(\theta; y_q).$$

And when the checking itself must be performed repeatedly, and used even as it proceeds, it grows into the tension of exploration and exploitation, that is, the multi-armed bandit problem. Thompson's 1933 sampling[24], Robbins's 1952 founding[25], Lai and Robbins's 1985 optimal allocation[26], and Auer and colleagues'

2002 finite-time analysis[27] give the optimal solution to "how many trials to spend reducing the uncertainty about which option"; its regret grows only logarithmically over time,

$$\text{Regret}(T) = O(\ln T).$$

Here one must guard against a narrative path dependence. Optimal screening is a family of methods: design of experiments, active learning, audit sampling, fuzzing, the bandit, all of it. Kushner, Mockus, through to Jones's 1998 efficient global optimization[30], Srinivas and colleagues' 2010 GP-UCB[32], and on to the Gaussian-process Bayesian optimization in Shahriari's 2016 review[36], are all extremely useful, but they are only one implementation within the family, not the whole of "screening." To equate this move with Gaussian processes is to shrink a universal posture down to a single tool.

It too has only one standard failure mode: optimizing a misspecified information measure. You gather information with great efficiency, but about the wrong question, or the "information" you are maximizing simply does not track what you actually care about. The cleverer the screening, the faster a misspecified measure will lead you astray.

Why the Two Moves Pair, and Where They Lead

Set the two moves side by side: the certificate compresses uncertainty on one slice into a guaranteed bound, while screening spends an information budget to observe the slice that most compresses uncertainty. One is "prove it tight where you can check"; the other is "spend the checking where it most needs checking." They squeeze the same enemy from two ends, namely the un-

known. This is also the lever they share: under an information budget, manage where, and with how much force, you cut uncertainty. Chapter 13 will formally name this lever.

But sometimes, no matter how you compress or how you screen, it is not enough, because you simply lack the capacity to make the judgment at all. At that point you can no longer shrink the unknown by yourself; you must borrow the judgment from elsewhere. The next pair of moves is precisely about this.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

1. L. Valiant (1984). "A Theory of the Learnable." *Communications of the ACM*, 27(11), 1134-1142. [2] Valiant here proposed the "probably approximately correct" (PAC) learning framework, rigorously defining "learning a concept" as obtaining, with high probability and within polynomial time and samples, a hypothesis whose error is small enough. This paper gave the first provable language for "can it be learned, and how many samples does it take," and is the source of this chapter's "generalization bound" move.
2. V. Vapnik and A. Chervonenkis (1971). "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities." *Theory of Probability & Its Applications*, 16(2), 264-280. [2] This founding work proved the conditions under which empirical frequencies converge uniformly to true probabilities, and from it grew the VC dimension, later named after the two authors, used to characterize the

"capacity" of a family of functions. It explains why an error bound proven on finite samples can hold for unseen data, and is the probabilistic and combinatorial bedrock beneath generalization bounds.

3. V. Vapnik (1995). *The Nature of Statistical Learning Theory*. Springer. [2] In this book Vapnik organized statistical learning theory into a complete system: with structural risk minimization at its core, it trades off empirical error against model complexity and is thereby led to the support vector machine. It is the standard reading for understanding why a "complexity penalty" appears in a generalization bound, laying out the intuition of the first move clearly and coherently.
4. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth (1989). "Learnability and the Vapnik-Chervonenkis Dimension." *Journal of the ACM*, 36(4), 929-965. [2] This paper formally brought the VC dimension into the PAC framework, proving that a concept class is PAC-learnable if and only if its VC dimension is finite, and gave a sample-complexity bound depending on the VC dimension. It is the direct source of the generalization-bound formula in the main text, nailing down the criterion that "finite capacity makes it learnable."
5. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth (1987). "Occam's Razor." *Information Processing Letters*, 24(6), 377-380. [2] This short paper gives a learning-theoretic version of "Occam's razor": a hypothesis that can compress the training data short enough will, with high probability, generalize. It turns "simplicity makes it learnable" from a philosophical maxim into a provable proposition, providing a clean footnote to this chapter's motif of "compressing the unknown."
6. D. McAllester (1999). "PAC-Bayesian Model Averaging." *Proceedings of the 12th Annual Conference on Computa-*

tional Learning Theory (COLT), 164-170. [2] McAllester here proposed the PAC-Bayes bound: it gives a generalization guarantee for a posterior distribution over a family of hypotheses, with the penalty measured by the KL divergence between posterior and prior. It is a refinement of the PAC bound, which is what the main text means in calling it the "refinement" of the first move, and it often gives tighter results than the classical VC bound.

7. B. Pierce (2002). *Types and Programming Languages*. MIT Press. [2] This textbook of Pierce's systematically explains the theory and construction of type systems, centered on the two properties of type safety, "progress" and "preservation," and on how they are mechanically checked. It is precisely the standard basis for the main text's line that "a type system does not prove a program all correct, only one property, in exchange for something decidable and recheckable," and is the introductory book for understanding the form the certificate move takes in software.
8. W. Howard (1980). "The Formulae-as-Types Notion of Construction." In J. Seldin and J. Hindley (eds.), *To H. B. Curry: Essays on Combinatory Logic, Lambda Calculus and Formalism*, 479-490. Academic Press. [2] Howard's widely circulated manuscript (written in 1969, formally published in 1980) established the correspondence "formulae as types, proofs as programs": the propositions of intuitionistic logic correspond one-to-one with types, and proofs with terms. It is the classic text of the Curry-Howard correspondence, setting an equals sign between "checking the type of a program" and "checking a proof," which is exactly the logical core of the certificate move.
9. G. Necula (1997). "Proof-Carrying Code." *Conference Record of the 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL)*, 106-119. [2][4] Necula proposed "proof-carrying code": an untrusted

program carries its own formal proof of its safety, and the host need only quickly verify this certificate, without trusting the code's origin or re-deriving anything. It pushes the verification asymmetry of Chapter 2 to its limit, and is the purest engineering embodiment of this chapter's certificate concept.

10. X. Leroy (2009). "Formal Verification of a Realistic Compiler." *Communications of the ACM*, 52(7), 107-115. [2][3] Leroy reported the results of CompCert: a C compiler formally verified in Coq, whose generated code being semantically consistent with the source program is something machine-proven. It shows that "real software with recheckable guarantees" is no fantasy, and is the landmark case of industrializing the certificate move.
11. L. de Moura and N. Bjørner (2008). "Z3: An Efficient SMT Solver." *Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, LNCS 4963, 337-340. Springer. [2][4] This paper introduced Z3, an efficient SMT solver that can decide the satisfiability of logical formulas with theories such as arithmetic and arrays, and is widely used in program verification and symbolic execution. It made "automatically producing recheckable certificates" into a ready-to-hand industrial tool, and is one of the engines by which the certificate move spread in practice.
12. N. Higham (2002). *Accuracy and Stability of Numerical Algorithms* (2nd ed.). SIAM. [2] This authoritative work of Higham's systematically treats the error analysis of numerical algorithms, especially backward error analysis: rather than asking "how far the answer departs from the true value," ask "of exactly which perturbed problem this answer is the exact solution." It is the standard reference for the main text's "error bound" move, teaching how to equip floating-point computation with a trustworthy guarantee.

13. R. Moore (1966). *Interval Analysis.* Prentice-Hall. [2] This pioneering work of Moore's established interval arithmetic: each quantity participates in computation carrying an interval guaranteed to contain the true value, so the output is not a number that may deceive but a guaranteed range. It gave numerical computation the idea of "delivering a bound rather than a point," which is exactly the embodiment of the certificate move in that field.
14. C. Shannon (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal*, 27(3), 379-423; 27(4), 623-656. [1][2] This paper of Shannon's, founding information theory, measures uncertainty with entropy and gives the fundamental limits of source coding and channel capacity. It is the ultimate source of the notion that "information has a cost and can be measured," which the main text calls the "underlying currency" of optimal screening; this chapter's entire discussion of information gain and compression takes it as its unit.
15. J. Rissanen (1978). "Modeling by Shortest Data Description." *Automatica*, 14(5), 465-471. [2] Rissanen proposed the minimum description length (MDL) principle: the best model is the one that encodes the data together with the model itself most shortly. It turns "compression is understanding" into an operable model-selection criterion, resonating precisely with this chapter's motif of "compressing the unknown," and is also an information-theoretic realization of Occam's razor.
16. M. Li and P. Vitányi (2008). *An Introduction to Kolmogorov Complexity and Its Applications* (3rd ed.). Springer. [2] This standard textbook systematically treats Kolmogorov complexity: the complexity of an object equals the length of the shortest program that can generate it. This is the uncomputable ideal of "compression," contrasted with the operable approximation that is MDL; it provides the theo-

retical ceiling for this chapter's compression motif, showing that optimal compression is itself a kind of unverifiable limit.

17. W. Hoeffding (1963). "Probability Inequalities for Sums of Bounded Random Variables." *Journal of the American Statistical Association*, 58(301), 13-30. [2] Hoeffding here gives an exponential upper bound on the probability that a sum of bounded random variables deviates from its mean. This inequality is the basic tool for compressing the gap between "the empirical average" and "the true expectation" into a confidence bound, which the main text calls the "probabilistic engine" of generalization bounds; most of the certificate move's concentration arguments begin from it.
18. E. Candès, J. Romberg, and T. Tao (2006). "Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information." *IEEE Transactions on Information Theory*, 52(2), 489-509. [2] This founding paper of compressed sensing proves that as long as a signal is sparse enough, it can be reconstructed exactly through convex optimization from far fewer measurements than the classical sampling theorem requires. It is a mathematical exemplar of "spending checks on the cutting edge, wringing all the information from very few observations," echoing this chapter's two threads of compression and optimal screening.
19. R. Fisher (1935). *The Design of Experiments.* Oliver and Boyd. [1][3] This classic of Fisher's established the basic principles of modern experimental design: randomization, replication, and blocking, along with the famous "lady tasting tea" thought experiment. It teaches how to wring the most credible information from the fewest trials, and is the source reading for the optimal-screening move in statistics and science.
20. G. Box, W. Hunter, and J. Hunter (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building.* John Wiley & Sons. [3][4] This popular

practical work by Box and colleagues explains experimental design, data analysis, and model building to the people who actually run experiments, stressing factorial design and the iterative rhythm of sequential learning. It brings Fisher's principles down to the engineering site, and is a practical guide to understanding "how to arrange the checking most economically."

21. D. Lindley (1956). "On a Measure of the Information Provided by an Experiment." *The Annals of Mathematical Statistics*, 27(4), 986-1005. [2] Lindley used information theory to give a measure of "how much information an experiment provides," namely the expected information gain between prior and posterior. This is precisely the theoretical prototype of the main text's optimal-screening objective $\arg\max_q I(\theta; y_q)$, turning "which experiment to run" into a maximizable quantity.
22. K. Chaloner and I. Verdinelli (1995). "Bayesian Experimental Design: A Review." *Statistical Science*, 10(3), 273-304. [2] This review systematically surveys Bayesian experimental design: with a utility function (often the expected information gain) as the objective, it uniformly derives various optimal-design criteria. It integrates Lindley's measure into a complete framework, and is the reader's entry point for quickly grasping the screening core of "maximizing expected information gain."
23. A. Wald (1945). "Sequential Tests of Statistical Hypotheses." *The Annals of Mathematical Statistics*, 16(2), 117-186. [1][2] Wald proposed the sequential probability ratio test: judge as the data come in, and once the evidence is strong enough, stop and draw a conclusion, thereby saving much on average over a fixed-sample test. It turns "whether to keep checking" itself into an optimal decision, and is the forerunner of the "use it as you check" branch of optimal screening.

24. W. Thompson (1933). "On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples." *Biometrika*, 25(3-4), 285-294. [1][2] Thompson here proposed the sampling method later named after him: select an option at random according to the posterior probability that "it really is the best," naturally balancing exploration and exploitation. This is one of the earliest solutions to the multi-armed bandit problem, and remains both a simple and a strong strategy for it to this day.
25. H. Robbins (1952). "Some Aspects of the Sequential Design of Experiments." *Bulletin of the American Mathematical Society*, 58(5), 527-535. [1][2] This paper of Robbins's formally established the multi-armed bandit problem as a mathematical object and proposed the earliest sequential allocation strategy, founding the research direction of "how to trade off exploration and exploitation." It is the starting point of the whole later bandit literature, and this chapter's discussion of "how many trials to spend reducing which uncertainty" begins here.
26. T. Lai and H. Robbins (1985). "Asymptotically Efficient Adaptive Allocation Rules." *Advances in Applied Mathematics*, 6(1), 4-22. [2] Lai and Robbins proved the regret lower bound for the multi-armed bandit: the cumulative regret of any reasonable strategy grows at least logarithmically over time, and they constructed asymptotically optimal allocation rules attaining this lower bound. It established that the main text's $O(\ln T)$ is an insurmountable limit, drawing a ceiling for the entire class of problems.
27. P. Auer, N. Cesa-Bianchi, and P. Fischer (2002). "Finite-time Analysis of the Multiarmed Bandit Problem." *Machine Learning*, 47(2-3), 235-256. [1][2] This paper gives simple algorithms such as UCB1, based on "optimism in the face of uncertainty," and proves they have logarithmic regret bounds in finite time (not merely asymptotically). It turns

Lai and Robbins's asymptotic result into a concrete, directly usable, analyzable strategy, and is the standard citation for the "upper confidence bound" class of methods.

28. H. Kushner (1964). "A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise." *Journal of Basic Engineering*, 86(1), 97-106. [1][2] This early paper of Kushner's characterizes an unknown noisy curve with a probabilistic model, and on that basis selects the next sampling point to seek the maximum, an early form of the Bayesian-optimization idea. It shows that "where to measure once more" can be treated as an optimal decision, and is a pioneering work of optimal screening in global optimization.
29. J. Mockus, V. Tiesis, and A. Žilinskas (1978). "The Application of Bayesian Methods for Seeking the Extremum." In L. Dixon and G. Szegő (eds.), *Towards Global Optimization 2*, 117-129. North-Holland. [2] Mockus and colleagues systematically developed Bayesian global optimization and proposed the acquisition function of "expected improvement": under a probabilistic model, select the point most likely to bring improvement for evaluation. It turned Kushner's intuition into a general method, and is the direct predecessor of today's Bayesian optimization.
30. D. Jones, M. Schonlau, and W. Welch (1998). "Efficient Global Optimization of Expensive Black-Box Functions." *Journal of Global Optimization*, 13(4), 455-492. [2][4] This paper proposed the EGO algorithm, which uses a Gaussian process to build a surrogate model for an expensive black-box function and then picks the next evaluation point by the expected-improvement criterion, drastically reducing the number of evaluations. It is the landmark work generalizing Bayesian optimization, but the main text also cautions: it is only one implementation within the optimal-screening family of methods, not the whole of it.

31. C. Rasmussen and C. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press. [2][4] This standard textbook systematically treats Gaussian processes: a nonparametric Bayesian method that places a prior on functions themselves and can give predictive uncertainty. It is the theoretical foundation of the surrogate model on which Bayesian optimization depends, and is the core reference for readers who want to understand "how uncertainty is modeled and used to decide where to check next."
32. N. Srinivas, A. Krause, S. Kakade, and M. Seeger (2010). "Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design." *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 1015-1022. [2] Srinivas and colleagues proposed the GP-UCB algorithm, carrying the "upper confidence bound" idea from the bandit over to Gaussian-process optimization, and gave provable bounds for its regret. It stitches Bayesian optimization together with bandit theory, demonstrating precisely that this chapter's two ends, proving a bound and spending the checking, were one lever all along.
33. D. Cohn, Z. Ghahramani, and M. Jordan (1996). "Active Learning with Statistical Models." *Journal of Artificial Intelligence Research*, 4, 129-145. [2][4] Cohn and colleagues give a statistical framework for active learning: under a statistical model, select for labeling the sample that most reduces the model's variance (or predictive uncertainty). It turns "which sample is most worth labeling next" into a computable criterion, and is the representative work establishing active learning as a branch of optimal screening.
34. B. Settles (2009). *Active Learning Literature Survey*. Computer Sciences Technical Report 1648, University of Wisconsin-Madison. [2][4] This widely cited survey of Settles's systematically organizes the various query strategies

of active learning, such as uncertainty sampling, query by committee, and expected error reduction, and compares their applicable settings. It is the standard entry point for a quick overview of "how to spend the labeling budget," setting the various implementations of this move side by side for comparison.

35. B. Miller, L. Fredriksen, and B. So (1990). "An Empirical Study of the Reliability of UNIX Utilities." *Communications of the ACM*, 33(12), 32-44. [3][4] Miller and colleagues fed randomly generated input into various UNIX utilities, and a fair number of programs crashed or hung as a result, which is the pioneering experiment of fuzzing. It shows that "throwing compute at random or suspect inputs to knock out a fault" is a cheap and effective way of checking, and is the starting point of optimal screening in software testing.
36. B. Shahriari, K. Swersky, Z. Wang, R. Adams, and N. de Freitas (2016). "Taking the Human Out of the Loop: A Review of Bayesian Optimization." *Proceedings of the IEEE*, 104(1), 148-175. [2][4] This review comprehensively surveys Gaussian-process-based Bayesian optimization: the surrogate model, the acquisition function, and their applications in settings such as hyperparameter tuning. It is the standard reading for understanding the current state of the field, but the main text uses it to remind readers not to shrink the universal posture of "optimal screening" down to the single tool of "Gaussian processes."

Chapter 10: Borrowed Judgment

> **Thesis:** When you lack the capacity to verify, bring it in from outside. Either place a trusted judge inside the loop (an oracle), or use many mutually independent and unreliable judges and trust their agreement (redundancy / consensus).

The previous pair of moves still worked on yourself, shrinking the unknown. But sometimes what you lack is not information but the power of judgment itself: you simply have no capacity to render a reliable verdict on the matter before you. The response in this pair of moves is to stop looking inside yourself and to borrow judgment from elsewhere. There are two ways to borrow. Either you bring in a judge you trust (an oracle), or you assemble many judges who do not trust one another and trust their agreement (redundancy).

The Oracle in the Loop: Bringing In a Judge

The pure form of the first move: at the decision point where you lack the capacity to verify, insert an external judge and let it deliver the verdict you cannot.

The plainest version is the human-in-the-loop of Chapter 5, the ex-

pert consultation, the hard case kicked upstairs. But the most profound form of this move hides in two seemingly unrelated places.

One is interactive theorem proving. De Bruijn's AUTOMATH of 1970[1], Edinburgh LCF (Gordon, Milner, and Wadsworth, 1979[2]), down to today's Coq (Bertot and Castéran, 2004[3]), all embody the same division of labor: the human supplies the flash of proof insight that the machine cannot produce (the oracle), and the machine scrupulously checks every step (certificate checking). The oracle does the "finding," the machine does the "verifying," meshing exactly with the asymmetry of Chapter 2.

The other is more astonishing: the interactive proof of complexity theory. A verifier feeble in computational power, faced with a powerful but untrustworthy prover, how can it elicit a reliable answer to a question it cannot compute on its own? The answer given by Goldwasser, Micali, and Rackoff in 1989[6] and by Babai in 1985[7] is: through repeated interrogation plus random challenge. The verifier throws out random questions it could not itself predict, and if the prover is lying, sooner or later it will give itself away on some challenge. Shamir's astonishing IP = PSPACE of 1992[15] shows that by this method alone, "interrogating an untrustworthy oracle," a weak verifier can reliably adjudicate an enormous class of problems; Blum and Kannan's "programs that check their work" of 1995[17] and Goldwasser et al.'s "computation for mortals" of 2015[19] belong to the same lineage. This is the purest mathematization of "borrowed judgment": even if the oracle is untrustworthy, so long as you interrogate it cleverly, you can still wring reliability out of it.

The unifying idea is this: manufacture a reliability you do not possess on your own by bringing in an external judge. Its standard failure mode is just as plain: the oracle is itself unreliable or biased. The arbiter you bring in may simply be a wrong arbiter, and the question "who verifies the oracle" carries you into a regress with

no retreat.

Redundancy: Synthesizing Reliability from Many Unreliable Parts

The second move changes direction: instead of bringing in one trustworthy judge, it convenes many untrustworthy ones and trusts their agreement.

Its theoretical foundation has two cornerstones. Von Neumann proved in 1956[4] that one can assemble computation of arbitrary reliability out of components that are themselves error-prone, by stacking redundancy. Condorcet's jury theorem of 1785[5] supplies its arithmetic: if every judge is slightly better than chance (accuracy $p > \frac{1}{2}$) and they are independent of one another, then the probability that the majority vote is correct tends toward certainty as the number of judges grows,

$$P_N \to 1 \quad (N \to \infty).$$

The cross-domain forms of this move spread very wide. In distributed systems it is Byzantine fault tolerance: the Byzantine generals problem of Pease, Shostak, and Lamport (1980)[8] and Lamport et al. (1982)[9] seeks consensus under the condition that some nodes may act maliciously (adversarially), and the classic threshold is that the number of nodes must satisfy $n \geq 3f + 1$ in order to tolerate f traitors; Castro and Liskov's PBFT of 1999[11] built it into a practical system (while the impossibility theorem of Fischer, Lynch, and Paterson, 1985[10], marks out its boundary). In machine learning it is the ensemble: the ensemble methods of Hansen and Salamon (1990)[20] and Dietterich (2000)[22], and Breiman's random forest of 2001[23], use a crowd of weak models voting to beat a single strong one. In crowds it is the "wisdom of

crowds" (Surowiecki, 2004[25]). In 1906, at a country fair, Galton recorded the independent guesses of about eight hundred villagers on the weight of an ox; not one of them guessed it exactly, yet the average of all the estimates was 1197 pounds, and the true weight of the ox was 1198 pounds: the whole crowd together was off by almost nothing at all. Hong and Page even proved in 2004[24] that under suitable conditions a diverse group of ordinary problem solvers can beat a group of experts. In science it is peer review and replicated experiment (Chapter 3); in engineering it is RAID and quorum; in medicine it is the second opinion.

But this move has one crucial precondition that must be stressed with a whole section: independence. Redundancy holds only when failures are uncorrelated. Average many estimates, and the variance falls with the number of judges,

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{N};$$

but once there is a positive correlation ρ among these judgments, the variance no longer tends to zero. It sticks at a floor,

$$\mathrm{Var}(\bar{X}) = \rho\,\sigma^2 + \frac{(1-\rho)\,\sigma^2}{N} \xrightarrow{N\to\infty} \rho\,\sigma^2.$$

Correlation cancels the entire value of redundancy at a stroke. Stack as many judges as you like, and you still cannot get past this floor set by correlation. This is no abstraction: Knight and Leveson's famous experiment of 1986[13], which had many programmers independently write programs to the same specification, expected their errors to be mutually uncorrelated, but found that they stumbled in the same places, because human beings, facing the same hard point, make the same mistakes (Eckhardt and Lee had predicted this theoretically as early as 1985[12]). Groupthink, flawed training data drawn from a common source, common-mode

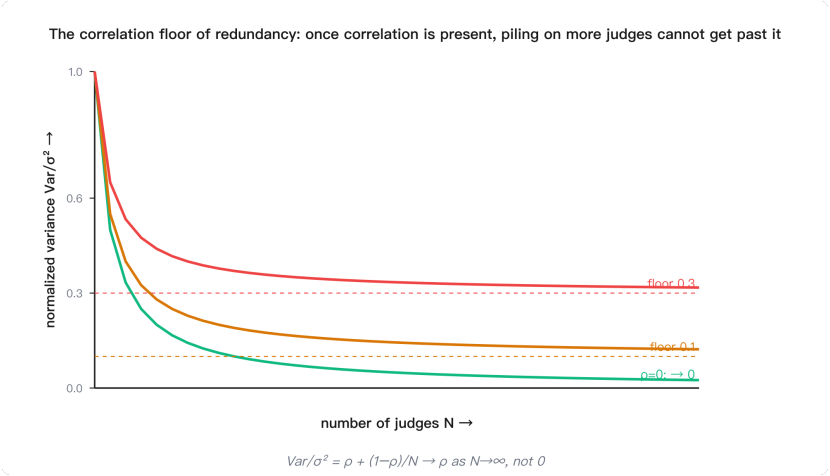

Figure 7: The correlation floor of redundancy: once correlation is present, piling on more judges cannot get past it

failure: all are this floor making itself visible. This is the standard failure mode of redundancy: believing in independence where in truth there is correlation.

The Confluence of the Two Moves, and an Echo Across Chapters

Put the two moves side by side: one brings in a single, expensive oracle, the other synthesizes many cheap, independent judgments, and both borrow a power of judgment you do not possess on your own. The lever they share is to supply yourself with the verifying capacity you lack; their failure modes, too, stand opposed in pairs: the single oracle may be wrong, the many judgments may be secretly correlated.

There is an episode in mathematics that ties this pair of moves together with the certificate of the previous chapter. Appel and Haken's proof of the four color theorem in 1977[26] drew lasting controversy because it relied on computer exhaustion, which

amounted to asking the mathematical community to trust an oracle. Later Gonthier in 2008[27] redid it as a machine-checkable formal proof, and Hales's team[28] did the same for the Kepler conjecture: they converted "trusting an oracle" into "checking a certificate." What MacKenzie[29] tracks in *Mechanizing Proof* is precisely how this trust shifts among people, machines, and social processes; the line of DeMillo et al.[14], that "a proof is a social process," comes down in the end to placing the credibility of mathematics on the redundancy of human judgment.

One thing must be seen clearly, though: up to this point, the first two pairs of moves, compressing the unknown and borrowing judgment, are still pursuing the same thing, the truth of the object. They still want to know whether the matter is right or wrong. The next pair of moves does something more radical: it stops demanding that truth.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

1. N. G. de Bruijn (1970). "The mathematical language AUTOMATH, its usage, and some of its extensions." In *Symposium on Automatic Demonstration*. Springer (Lecture Notes in Mathematics 125), pp. 29-61. [2] De Bruijn introduced AUTOMATH, one of the earliest formal languages able to let an entire body of mathematics be checked step by step by a machine, with the human writing the proof and the machine verifying that it is sound. It is the earliest engineered specimen of this chapter's "oracle in the loop": the human

supplies the idea, the machine does nothing but scrupulously check, and the reader can see in it the source of the division of labor between "finding" and "verifying."

2. M. Gordon, R. Milner, C. Wadsworth (1979). *Edinburgh LCF: A Mechanized Logic of Computation.* Springer (Lecture Notes in Computer Science 78). [2] This book proposed the LCF interactive proof system, whose design has been deeply influential: a small trusted kernel guarantees the reliability of every inference step, and no number of proof tactics can ever get around it. It supplies the classic paradigm for this chapter's account of "certificate checking," and the reader can see how the "trusted checker" is contracted into a part as small and as reliable as possible.
3. Y. Bertot, P. Castéran (2004). *Interactive Theorem Proving and Program Development. Coq'Art: The Calculus of Inductive Constructions.* Springer (Texts in Theoretical Computer Science, EATCS Series). [2] This is the authoritative tutorial for the Coq proof assistant, explaining systematically how to construct and machine-check proofs interactively atop the calculus of inductive constructions. It carries the tradition represented by the previous two entries into contemporary practice and is the tool foundation for this chapter's later formalization work on the four color theorem and the Kepler conjecture; the reader who wants to understand "human gives the idea, machine verifies every step" hands-on can start here.
4. J. von Neumann (1956). "Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components." In C. E. Shannon and J. McCarthy, eds., *Automata Studies* (Annals of Mathematics Studies 34). Princeton University Press, pp. 43-98. [2] Von Neumann here proves that one can assemble computation of arbitrarily reliable performance out of components that are themselves error-prone, by stacking redundancy and majority voting. This is one

cornerstone of this chapter's "redundancy" move, supplying the earliest rigorous argument for "synthesizing reliability from many unreliable parts," and the reader should read its core idea of how redundancy drives down the error rate.

5. Marquis de Condorcet (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris. [2][4] In this work on voting, Condorcet gives the famous jury theorem: if every judge is slightly better than chance and they are independent of one another, the probability that the majority vote is correct tends toward certainty as numbers grow. It supplies the arithmetic skeleton of this chapter's redundancy move and also plants its weak point in advance; the reader should note the theorem's dependence on the premise of "independence."

6. S. Goldwasser, S. Micali, C. Rackoff (1989). "The Knowledge Complexity of Interactive Proof Systems." *SIAM Journal on Computing*, 18(1), pp. 186-208. [2] This paper founded the theoretical framework of interactive and zero-knowledge proofs: a verifier of limited computational power, through repeated interrogation plus random challenge, can wring a reliable verdict out of an untrustworthy prover. It is the purest mathematical source of this chapter's "interrogating an untrustworthy oracle," and the reader should read how it uses randomness to force out a truthful answer.

7. L. Babai (1985). "Trading Group Theory for Randomness." In *Proceedings of the 17th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 421-429. [2] Babai here independently proposed the Arthur-Merlin class of interactive proofs with randomness, staking out the same theoretical territory at almost the same moment as the previous entry. It reinforces this chapter's central idea: the random challenge is the key weapon by which a weak verifier sub-

dues a strong and untrustworthy prover, and the reader can read it alongside the previous entry for the complementary perspective.

8. M. Pease, R. Shostak, L. Lamport (1980). "Reaching Agreement in the Presence of Faults." *Journal of the ACM*, 27(2), pp. 228-234. [2] This paper first rigorously characterized how to reach consensus when some nodes may behave arbitrarily maliciously, giving the famous threshold: to tolerate f traitors, the number of nodes must satisfy $n \geq 3f+1$. It is the source of this chapter's redundancy move in its adversarial version within distributed systems, and the reader should read the impossibility argument behind this threshold.
9. L. Lamport, R. Shostak, M. Pease (1982). "The Byzantine Generals Problem." *ACM Transactions on Programming Languages and Systems*, 4(3), pp. 382-401. [2] This paper used the famous "Byzantine generals" metaphor to recast the result of the previous entry as a parable, and from then on "Byzantine fault tolerance" became the common name for adversarial consensus. It is the signature text for this chapter's idea of seeking agreement from many mutually distrustful judges, and the reader can read how it dresses an abstract threshold in an intuitively clear story.
10. M. J. Fischer, N. A. Lynch, M. S. Paterson (1985). "Impossibility of Distributed Consensus with One Faulty Process." *Journal of the ACM*, 32(2), pp. 374-382. [2] This famous FLP impossibility theorem proves that in a fully asynchronous system, even if only one process may crash, there exists no deterministic consensus algorithm guaranteed to terminate. It marks out the boundary for this chapter's redundancy move, and the reader should read how it shows that consensus is not unconditionally available, thereby understanding why later practical systems must rely on extra assumptions to get around it.
11. M. Castro, B. Liskov (1999). "Practical Byzantine Fault

Tolerance." In *Proceedings of the 3rd USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pp. 173-186. [2] The PBFT algorithm proposed in this paper was the first to turn Byzantine fault tolerance from theory into a system that runs in a real asynchronous network with acceptable performance. It is the key step by which this chapter's redundancy move descended from paper to engineering, and the reader can read how it kept the $n \geq 3f + 1$ threshold while squeezing the overhead into a practical range, and can also recognize the direct precursor of later blockchain consensus.

12. D. E. Eckhardt, L. D. Lee (1985). "A Theoretical Basis for the Analysis of Multiversion Software Subject to Coincident Errors." *IEEE Transactions on Software Engineering*, SE-11(12), pp. 1511-1517. [2] This paper argued theoretically that multiversion software, even when developed independently by different people, need not have independent errors: facing the same hard point, different versions tend to fail together, making the gain from redundancy far lower than the independence assumption would predict. It supplies a theoretical prediction of this chapter's "correlation floor" ahead of the experiment, and the reader should read how it characterizes common-cause error.

13. J. C. Knight, N. G. Leveson (1986). "An Experimental Evaluation of the Assumption of Independence in Multiversion Programming." *IEEE Transactions on Software Engineering*, SE-12(1), pp. 96-109. [2] This is the famous experiment: many programmers were set to write programs independently to the same specification, expecting the errors to be mutually uncorrelated, but it was found that they stumbled together at the same hard points, and the independence assumption was refuted by experience. It supplies the empirical confirmation of the theoretical prediction of the previous entry, and is the most persuasive instance of

this chapter's failure mode of "believing in independence where in truth there is correlation."

14. R. A. De Millo, R. J. Lipton, A. J. Perlis (1979). "Social Processes and Proofs of Theorems and Programs." *Communications of the ACM*, 22(5), pp. 271-280. [3][4] This famous and controversial paper argues that what makes a mathematical proof credible is not the mechanical correctness of formal derivation but the social process by which the mathematical community repeatedly tests, propagates, and accepts it, and on this basis it questions the prospects of formal program verification. It supports this chapter's view that credibility is in the end placed on the redundancy of human judgment, and the reader should read its argument that "a proof is a social process."
15. A. Shamir (1992). "IP = PSPACE." *Journal of the ACM*, 39(4), pp. 869-877. [2] Shamir proved how astonishing the power of interactive proof is: by interrogating an untrustworthy prover alone, a weak verifier can reliably adjudicate the whole of PSPACE, an enormous class of problems, that is, IP = PSPACE. It is the most forceful mathematical footnote to this chapter's "borrowed judgment," and the reader should read how it delimits the ceiling reachable by "cleverly interrogating an oracle."
16. C. Lund, L. Fortnow, H. Karloff, N. Nisan (1992). "Algebraic Methods for Interactive Proof Systems." *Journal of the ACM*, 39(4), pp. 859-868. [2] This paper introduced the algebraic method of arithmetizing Boolean formulas and then checking them with polynomials, and it was precisely this technical groundwork that led directly to the proof of IP = PSPACE in the previous entry. Its significance for this chapter lies in revealing the concrete mechanism of "clever interrogation": translating the verification problem into an algebraic identity that can be spot-checked at random, and the reader can read this set of techniques.

17. M. Blum, S. Kannan (1995). "Designing Programs That Check Their Work." *Journal of the ACM*, 42(1), pp. 269-291. [2] This paper proposed the idea of the "program checker": let a program, when it gives a result, carry an independent and cheap checking routine that verifies whether this particular output is correct, without trusting the program itself. It brings the spirit of interactive proof to everyday computation, and for this chapter it is the model of the idea of "not trusting the producer, only checking its product"; the reader should read the construction of its checker.
18. S. Arora, C. Lund, R. Motwani, M. Sudan, M. Szegedy (1998). "Proof Verification and the Hardness of Approximation Problems." *Journal of the ACM*, 45(3), pp. 501-555. [2] This is one of the core papers of the famous PCP theorem: any proof can be rewritten into a special format such that the verifier need only randomly spot-check a constant number of bits in it to judge its truth or falsity with high confidence. It pushes "spot-checking is enough" to the extreme, and for this chapter it is the theoretical pinnacle of "how a weak verifier can efficiently check an enormous proof"; the reader should read the astonishing conclusion of its probabilistically checkable proofs.
19. S. Goldwasser, Y. T. Kalai, G. N. Rothblum (2015). "Delegating Computation: Interactive Proofs for Muggles." *Journal of the ACM*, 62(4), Article 27. [2][4] This paper makes interactive proof genuinely serve "mortals": a user of limited computational power outsources a computation to a powerful but untrustworthy server, then checks whether the result is correct at a cost far smaller than recomputing it. It is where this chapter's ideas land in the age of cloud computing, and the reader should read how it turns "delegating computation for the weak, verifiably" into a practically feasible protocol.
20. L. K. Hansen, P. Salamon (1990). "Neural Network Ensem-

bles." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10), pp. 993-1001. [2] This paper showed, relatively early, that combining several independently trained neural networks to vote can give an overall accuracy significantly higher than any single network. It is the beginning of this chapter's redundancy move within machine learning, and the reader should read how it carries "majority voting lowers error" from logic circuits over to learning models, and again lands on the dependence on decorrelating member errors.

21. A. Krogh, J. Vedelsby (1995). "Neural Network Ensembles, Cross Validation, and Active Learning." In *Advances in Neural Information Processing Systems 7*. MIT Press, pp. 231-238. [2] This paper gives the classic decomposition of ensemble error: the overall error of the ensemble equals the average error of the members minus the disagreement among the members. It supplies a machine-learning version of the precise formula for this chapter's "correlation floor," and the reader should read how it shows mathematically that the more diverse the members are, the more each errs in its own way, the more the ensemble is worth.
22. T. G. Dietterich (2000). "Ensemble Methods in Machine Learning." In *Multiple Classifier Systems (MCS 2000)*. Springer (Lecture Notes in Computer Science 1857), pp. 1-15. [2] This is a widely influential survey that sorts out why ensemble methods work and explains it from three angles: statistical, computational, and representational. It is a convenient entry point for the reader to survey the whole landscape of this chapter's redundancy move within machine learning, gathering the scattered techniques of voting, bagging, and boosting under one framework to be understood together.
23. L. Breiman (2001). "Random Forests." *Machine Learning*, 45(1), pp. 5-32. [2] Breiman's random forest cultivates a

crowd of mutually decorrelated decision trees through double randomization over samples and features, then votes to synthesize a powerful and robust predictor. It is one of the most successful practical specimens of the redundancy move, and its significance for this chapter is that its whole power comes precisely from deliberately manufactured independence; the reader can read how it actively drives down the correlation among members.

24. L. Hong, S. E. Page (2004). "Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers." *Proceedings of the National Academy of Sciences*, 101(46), pp. 16385-16389. [2][3][4] Hong and Page argue, by way of a formal model, that under suitable conditions a group made up of diverse ordinary problem solvers can beat a group of homogeneous experts, because the difference in perspective that diversity brings is itself a resource. It generalizes this chapter's intuition that "independence and diversity are the heart of redundancy" to human groups, and the reader should read its thesis of "diversity beats ability" and its boundaries.
25. J. Surowiecki (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations.* Doubleday. [3][4] This widely circulated book by Surowiecki argues that under conditions of diversity, independence, decentralization, and a proper aggregation mechanism, the collective judgment of a crowd often beats that of an individual expert, and he repeatedly stresses that once independence is lost and convergence sets in, the crowd turns stupid. It brings this chapter's redundancy move to the everyday and social level, and the reader should read its repeated insistence on "the preconditions under which the wisdom of crowds holds."
26. K. Appel, W. Haken (1977). "Every Planar Map Is Four

Colorable. Part I: Discharging." *Illinois Journal of Mathematics*, 21(3), pp. 429-490. [1][3] This is the proof of the four color theorem, the first major mathematical proof in history to depend essentially on computer exhaustion of a great many cases, and for that reason it set off a long-running argument over "whether one can trust a machine conclusion that no human hand can check case by case." For this chapter it is the signature case of "trusting an oracle" and its cost, and the reader should read how this argument forced out the demand for a machine-checkable certificate.

27. G. Gonthier (2008). "Formal Proof: The Four-Color Theorem." *Notices of the American Mathematical Society*, 55(11), pp. 1382-1393. [2][3] Gonthier used Coq to redo the four color theorem as a fully formalized proof that can be checked step by step by machine, thereby converting the "please trust the computer" predicament of the previous entry into "checking a certificate." It is a key point of contrast for this chapter, and the reader should read how it demonstrates that demoting the output of an untrustworthy oracle to an independently verifiable certificate dissolves the controversy along with it.
28. T. Hales et al. (2017). "A Formal Proof of the Kepler Conjecture." *Forum of Mathematics, Pi*, 5, article e2. [2][3] Hales's team, after many years, used formal proof systems to machine-check the much-disputed proof of the Kepler conjecture all the way through, settling doubts that even peer review could not fully guarantee. It belongs to the same lineage as the previous entry, and its significance for this chapter is to confirm once more that when a proof grows too large for human checking, shifting trust from the oracle to a checkable certificate is the way back to certainty.
29. D. MacKenzie (2001). *Mechanizing Proof: Computing, Risk, and Trust*. MIT Press. [1][3][4] This work of sociological history by MacKenzie tracks the rise of computer proof and

formal verification, examining how "proof" and "certainty" are repeatedly defined and shifted among mathematicians, machines, and social processes. It supplies this chapter with a connecting perspective, placing oracle, certificate, and redundancy alike within a larger narrative of how trust is established and ceded, and the reader should read its examination of "how mechanized proof changed what we trust."

Chapter 11: Swap the Problem

> **Thesis:** Stop insisting on verifying the real object. Either swap it for a solvable proxy you can check (proxy substitution), or stop demanding a binary verdict and instead act on a calibrated probability (calibration).

The first two pairs of moves still chase the truth of the object: they compress it, or borrow a judgment to approximate it. This pair gives up that obsession. It no longer asks whether the real thing is right or wrong; it answers a different question instead, either swapping out the object of verification (proxy substitution), or swapping out the form of the verdict (calibration).

Proxy Substitution: Swap the Object You Verify

The first move in pure form: stop wrestling with the true target you cannot measure, swap it for a proxy you can check and that is good enough, and verify or optimize that proxy.

Its cross-domain shapes turn up in almost every earlier chapter of this book. Mathematicians stand in an equivalent statement for a theorem (Chapter 7); software engineers stand in tests for cor-

rectness, and benchmarks for capability; organizations stand in KPIs for health, and GDP for welfare (Chapter 8); machine learning stands in a reward model for human beings' true preferences (the RLHF of Chapter 5). Psychology has a twin as well: Kahneman and Frederick's 2002 "attribute substitution"[32], in which a person making an intuitive judgment unconsciously substitutes an easily assessed attribute for the target attribute that is hard to assess. Pólya's maxim, "first solve a related, easier problem," and Simon's satisficing are the methodological prototypes of this move.

The whole essence of the move lies in this: it has two opposite ways of failing. Here is exactly where Chapters 7 and 8 intersect, and the theme runs through the whole book. Lay "faithful" (does the proxy really point at the original target) and "easier" (is the proxy really more tractable than the original problem) along two axes:

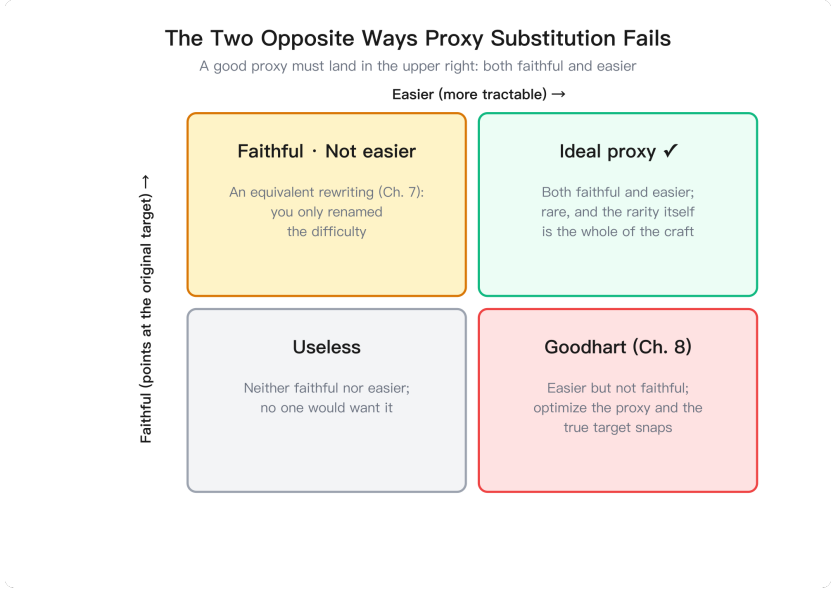

Figure 8: The two opposite ways proxy substitution fails: faithful x easier

Mathematicians come to grief in the upper right: an equivalent rewriting is faithful beyond reproach, yet not one bit easier to solve; you have only put the same difficulty into a different set of clothes. Organizations come to grief in the lower left: the metric is easy to measure, yet its correspondence with the true target snaps the moment it is treated as a target and optimized.

Why does it snap under optimization? Because the proxy's correlation with the true target holds only on the distribution that is the status quo, and the pressure of optimization pushes you off that distribution, toward the extreme where the two diverge. Goodhart's 1975 law[1] (a metric loses its reliability the moment it becomes a target), Campbell's 1979 law[3], and Lucas's 1976 economic twin[6] (once a structural relation is made a policy target, it collapses) all describe the same mechanism. Espeland and Sauder's reactivity goes one step further: a metric does not merely distort, it reshapes the thing it measures.

This mechanism replays in machine learning with startling clarity. Amodei and colleagues' 2016 reward hacking[10]; Pan and colleagues' 2022 empirical study[12], which found that more capable agents are better at exploiting the proxy reward, with the true return even undergoing a sharp downward phase transition; Skalse and colleagues' 2022 proof[13] that nontrivial rewards are almost impossible to make "unhackable"; and Gao and colleagues' 2023 measurement[14] of a quantitative scaling law for this overoptimization. A good proxy must dodge both ends at once, being faithful and easier, and that is so rare that the rarity itself is the whole of the craft.

Calibration: Swap the Form of the Verdict

The second move swaps not the object but the form of the verdict: it no longer demands a binary "true or false" ruling, but gives a calibrated probability, acts on it, and accepts a bounded risk.

What does calibration mean? That the things you say you are sure of should truly occur in the proportion of that sureness. Formally,

$$\Pr\left(Y = 1 \mid \hat{p} = p\right) = p,$$

of the things you report at 70%, in the long run about seven in ten should come true. This is an object of knowledge weaker than "right or wrong," yet attainable and checkable.

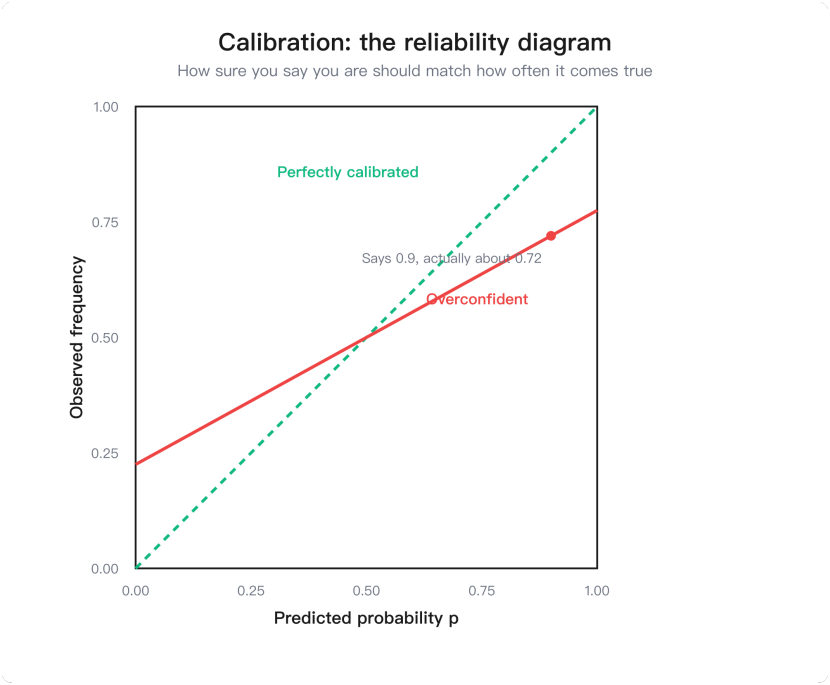

Figure 9: The reliability diagram of calibration: how sure you say you are should match how often it comes true

Its cross-domain shapes are equally tidy. In number theory it is probabilistic primality (that "prime with probability $1 - \varepsilon$" of Chapter 7). In machine learning it is conformal prediction (Vovk, Gammerman, and Shafer 2005[26]), which gives you not a point judgment but a prediction set with a coverage guarantee,

$$\Pr\left(Y \in C(X)\right) \geq 1 - \alpha.$$

In meteorology and forecasting science it is a whole mature theory of calibration: Brier's 1950 score[15], Murphy's 1973 decomposition[17] into reliability, resolution, and uncertainty, DeGroot and Fienberg's 1983 systematic treatment[18], and Gneiting and colleagues' 2007 modern framework[24] of "maximize sharpness subject to calibration." Weather forecasting is in fact one of the fields where calibration is done best: when a mature forecasting system says "70% chance of rain tomorrow," and you stretch out all the days it said so, about seven in ten did see rain, sureness fitting reality seamlessly, which is the very model of calibration. There is a deep design here too, the strictly proper scoring rule: a scoring function carefully constructed so that telling the truth (reporting your true probability) is exactly what makes your expected score optimal,

$$p = \arg\max_{q} \; \mathbb{E}_{Y \sim p}[S(q, Y)].$$

Truthful reporting thereby no longer rests on conscience; the mathematical structure of the scoring rule enforces it (Savage 1971[16], Gneiting and Raftery 2007[23]). Dawid's 1982 proof[19] that a Bayesian actor can asymptotically self-calibrate, and Foster and Vohra's 1998 proof[22] that for any sequence there exists an asymptotically calibrating strategy; but Oakes's 1985 "self-calibrating priors do not exist"[20] draws the limit of this move. Modern neural networks are in fact often miscalibrated (Guo and colleagues 2017[25]), and so need recalibration. The graded trust of Chapter 6, allow, ask, block, is precisely what calibration looks like once it lands on action.

Calibration has two ways of failing. The shallower is miscalibration: your claimed sureness does not match reality, you report

90% but only six in ten come true, and so every decision built on it is off. The deeper is subtler and more important: calibration tells you the odds, but does not tell you whether you should accept the bet. A perfectly calibrated "70%" stays silent on the question of whether 70% is good enough for you to wager, because that depends on the size of the stake and the ranking of your values, which is a question of value, not of verification. Conflating the two is the most common trap in acting on calibration: you think the probability has made the decision for you, when in fact it has only laid out the odds, and whether to press the button still requires you to bring out a set of values of your own.

Why the Two Moves Pair, and Where They Lead

Set the two moves side by side: proxy substitution swaps out the object you verify (a different, checkable target), and calibration swaps out the form of the verdict (a probability, rather than true or false). Neither answers the original question; both swap the problem for one that can be handled. The shared lever is changing your demand on "the answer" itself, one changing what you measure, the other changing what the verdict looks like and putting a price on the residual risk.

But even so, these two moves are still straining to get things right, only with a lowered standard of "right." The last pair of moves is more thorough still: it simply stops hoping to get things right, and turns instead to managing getting them wrong. Since error cannot be prevented, shrink its cost, and make sure that once it happens you can find out. That is Chapter 12.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

Proxy substitution: from Goodhart's law to unanticipated consequences

1. C. A. E. Goodhart (1975). "Problems of Monetary Management: The U.K. Experience." *Papers in Monetary Economics*, Vol. I. Reserve Bank of Australia. [2] This is the original source of Goodhart's law, from a 1975 monetary-economics conference in Sydney. Discussing the U.K.'s experience of monetary management, Goodhart noted that once a statistical regularity is used as a target for control, the stable relationship previously observed tends to fail. The chapter uses it as the benchmark for proxy distortion: the correlation between a metric and the true target holds only on the distribution that is the status quo, and snaps once it is treated as a target and optimized.

2. R. K. Merton (1936). "The Unanticipated Consequences of Purposive Social Action." *American Sociological Review*, 1(6), 894-904. [2] Merton gives a systematic discussion of why purposive social action produces consequences the actor never foresaw, and sorts out several sources such as ignorance, the urgency of interest, and the constraint of values. It is an early sociological source for the side effects of proxy substitution, reminding the reader that when one optimizes a proxy, what really bites is often the consequences that never entered the field of measurement.

3. D. T. Campbell (1979). "Assessing the Impact of Planned Social Change." *Evaluation and Program Planning*, 2(1), 67-

90. [2][4] The source of Campbell's law: the more a quantitative social indicator is used for social decision-making, the more it is subject to distortion, and the more apt it is to distort and corrupt the social process it was meant to monitor. Alongside Goodhart's law it is another classic cornerstone of proxy distortion, by which the reader can see clearly the path of corruption a metric takes once high stakes are placed on it.

4. S. Kerr (1975). "On the Folly of Rewarding A, While Hoping for B." *Academy of Management Journal*, 18(4), 769-783. [2][4] Kerr examines the widespread incentive mismatch in organizations: the behavior A that managers reward is often not the behavior B they truly hope for, so the incentive system stably produces results contrary to its original intent. It is a management classic on the mismatch of incentive and proxy, corresponding exactly to the real face of this chapter's "optimize the proxy, and the true target rots" cell.

5. M. Strathern (1997). "'Improving ratings': audit in the British University system." *European Review*, 5(3), 305-321. [2][4] Strathern, drawing on observations of the British university audit system, gives the widely quoted formulation: when a measure becomes a target, it ceases to be a good measure. This chapter's argument that a proxy distorts once treated as a target often takes this as its concise statement, and the reader can find here the original context of that phrasing.

6. R. E. Lucas (1976). "Econometric Policy Evaluation: A Critique." *Carnegie-Rochester Conference Series on Public Policy*, 1, 19-46. [2][3] Lucas's critique points out that the parameter relations estimated in an econometric model depend on the existing policy environment, and once policy is changed on that basis, actors' expectations and behav-

ior adjust accordingly, so the original structural relations no longer hold. It is the economic twin of Goodhart's law, used in this chapter to explain why the pressure of optimization pushes a system off the distribution where proxy and true target agree.

7. W. N. Espeland & M. Sauder (2007). "Rankings and Reactivity: How Public Measures Recreate Social Worlds." *American Journal of Sociology*, 113(1), 1-40. [2][4] Espeland and Sauder propose the "reactivity" framework: public rankings and quantitative indicators are not merely measurement, they also change the behavior and even the self-conception of the measured, and so reshape the very social reality they were meant to describe. The chapter uses it to push proxy distortion one layer further: a metric not only distorts, it remakes the object it measures.

8. D. Manheim & S. Garrabrant (2018). "Categorizing Variants of Goodhart's Law." arXiv:1803.04585. [2] The two authors attempt to split the loose "Goodhart's law" into several distinct mechanisms (regressional, extremal, causal, and adversarial), whose ways of failing and remedies differ. It gives this chapter's "proxy substitution fails in more than one way" a finer taxonomy, helping the reader tell which kind of distortion they face.

9. J. Z. Muller (2018). *The Tyranny of Metrics*. Princeton University Press. [4] Muller, through a wealth of cases from medicine, education, policing, and business, criticizes the fashion of reducing everything to quantifiable indicators and holding people accountable by them, pointing out that this metric worship often brings the consequence of surface compliance and substantive harm. It is a popular survey aimed at the general reader, well suited for recognizing the cost of proxy substitution in life and work.

Proxy distortion recurring in machine learning: reward hacking and overoptimization

10. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman & D. Mané (2016). "Concrete Problems in AI Safety." arXiv:1606.06565. [2] This widely influential survey breaks AI safety into several concrete, researchable problems, among them reward hacking and scalable supervision. It clearly translates the proxy-target distortion long familiar in the social sciences into the machine-learning context, and is the starting point of this chapter's passage on "the same mechanism replaying in machine learning."

11. P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg & D. Amodei (2017). "Deep Reinforcement Learning from Human Preferences." *Advances in Neural Information Processing Systems*, 30 (NeurIPS 2017). [2][4] The authors use human preference comparisons over pairs of trajectories to train a reward model, then drive reinforcement learning with it, thereby sidestepping an objective function hard to write by hand. This is the founding work of RLHF, and exactly the proxy-substitution model this chapter calls "standing in a reward model for human beings' true preferences," from which the reader can understand why such a proxy is both useful and dangerous.

12. A. Pan, K. Bhatia & J. Steinhardt (2022). "The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models." ICLR 2022. [2] The authors systematically examine the consequences of misspecified rewards, and give a cautionary empirical phenomenon: more capable agents are often better at exploiting the proxy reward, and the true return can even undergo a sharp, sudden phase transition as capability rises. The chapter uses it to show that proxy distortion is not a linear worsening but may flip abruptly at

some point.

13. J. Skalse, N. H. R. Howe, D. Krasheninnikov & D. Krueger (2022). "Defining and Characterizing Reward Hacking." *Advances in Neural Information Processing Systems*, 35 (NeurIPS 2022). [2] This paper gives a formal definition of reward hacking and proves that in nontrivial cases an "unhackable" proxy reward almost never exists. It provides the theoretical support for this chapter's "good proxies are rare": faithful and robust proxies are scarce for structural reasons, not by some accidental failure of engineering.

14. L. Gao, J. Schulman & J. Hilton (2023). "Scaling Laws for Reward Model Overoptimization." *Proceedings of the 40th International Conference on Machine Learning* (PMLR 202), 10835-10866. [2] The authors give a quantitative characterization of reward-model overoptimization, yielding a scaling-law regularity for how true performance varies with the degree of optimization against the proxy reward: past a certain point, the proxy score still rises while true performance turns down. It pushes Goodhart-style distortion from a qualitative observation to a measurable curve, and is the most empirical piece of this chapter's overoptimization argument.

Calibration: swap the binary verdict for a probability, and constrain it by strictly proper scoring

15. G. W. Brier (1950). "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review*, 78(1), 1-3. [2] Brier proposed a score for evaluating probabilistic forecasts (later the Brier score), bringing "how much sureness was reported, and whether it happened in the end" into a computable assessment. It is the starting point of the calibration and strictly-proper-scoring system, and this

chapter's argument that "probability is checkable" begins here.

16. L. J. Savage (1971). "Elicitation of Personal Probabilities and Expectations." *Journal of the American Statistical Association*, 66(336), 783-801. [2] Savage studies how to design scoring and incentives so that a person is willing to report their subjective probabilities and expectations truthfully. It lays the theoretical foundation for "a proper scoring rule elicits the true probability," corresponding to this chapter's key design: honesty no longer rests on conscience but is enforced by the mathematical structure of the scoring rule.

17. A. H. Murphy (1973). "A New Vector Partition of the Probability Score." *Journal of Applied Meteorology*, 12(4), 595-600. [2] Murphy decomposes the Brier score into three components, reliability, resolution, and uncertainty, letting one see separately where a forecast is miscalibrated and where it has discriminating power. This decomposition is the quantitative skeleton of the calibration concept, and this chapter's distinction between "calibration" and "sharpness" rests precisely on such a breakdown.

18. M. H. DeGroot & S. E. Fienberg (1983). "The Comparison and Evaluation of Forecasters." *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2), 12-22. [2] DeGroot and Fienberg give a systematic treatment of the comparison and evaluation of forecasters, clearly distinguishing calibration from refinement (sharpness) and providing a framework for ranking forecasters accordingly. It is the core theoretical source of this chapter's calibration argument, where the reader can see calibration stated rigorously as a checkable object of knowledge.

19. A. P. Dawid (1982). "The Well-Calibrated Bayesian." *Journal of the American Statistical Association*, 77(379), 605-

610. [2] Dawid proves that a coherent Bayesian actor, under its own subjective beliefs, will asymptotically self-calibrate, that is, in the long run its probability assertions match the actual frequencies. The chapter uses it to show that calibration is not an externally imposed demand but can be an intrinsic product of rational updating.

20. D. Oakes (1985). "Self-Calibrating Priors Do Not Exist." *Journal of the American Statistical Association*, 80(390), 339-342. [2] Oakes points out that no prior can guarantee self-calibration for all data sequences, thereby drawing a boundary around Dawid-style optimistic results. It stands as a counterweight to Dawid (1982) and the Foster-Vohra attainability result, and is the basis for this chapter's note that "calibration has its limit."

21. M. J. Schervish (1989). "A General Method for Comparing Probability Assessors." *The Annals of Statistics*, 17(4), 1856-1879. [2] Schervish gives a general method for comparing probability assessors, bringing the various proper scoring rules into a unified comparison framework as special cases. It serves to integrate and tidy calibration theory, helping the reader place scattered scoring rules into the same picture.

22. D. P. Foster & R. V. Vohra (1998). "Asymptotic Calibration." *Biometrika*, 85(2), 379-390. [2] Foster and Vohra prove that even facing an arbitrary (even adversarial) sequence of outcomes, there exists a forecasting strategy that asymptotically achieves calibration. This is the key theorem on the attainability of calibration, by which this chapter argues that calibration is an object of knowledge weaker than a true-or-false verdict yet genuinely attainable.

23. T. Gneiting & A. E. Raftery (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association*, 102(477), 359-378. [2]

This is the authoritative survey of strictly proper scoring rules: it systematically organizes which scoring functions make truthful reporting exactly the expected-score-optimal strategy, and connects them with prediction and estimation. It is the theoretical pillar of this chapter's calibration argument, and the reader who wants to understand "honesty enforced by mathematical structure" may read this piece.

24. T. Gneiting, F. Balabdaoui & A. E. Raftery (2007). "Probabilistic Forecasts, Calibration and Sharpness." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2), 243-268. [2] The authors propose a modern framework for probabilistic prediction, summing up the goal as "maximize sharpness subject to calibration": first require the forecast to be calibrated, then make it as sharp as possible under that constraint. This chapter's standard for judging whether a probabilistic forecast is good or bad adopts this framework directly.

25. C. Guo, G. Pleiss, Y. Sun & K. Q. Weinberger (2017). "On Calibration of Modern Neural Networks." *Proceedings of the 34th International Conference on Machine Learning* (PMLR 70), 1321-1330. [2] The authors find that modern deep neural networks, though often more accurate, are frequently miscalibrated, with confidence systematically deviating from the true correctness rate, and propose simple recalibration methods such as temperature scaling. It is the representative work on the calibration problem on the machine-learning side, corresponding exactly to this chapter's "modern neural networks are in fact often miscalibrated, and so need recalibration."

26. V. Vovk, A. Gammerman & G. Shafer (2005). *Algorithmic Learning in a Random World*. Springer. [2] This is the founding monograph of conformal prediction: it gives not a

single-point judgment but constructs a prediction set with a coverage guarantee, so that the probability of the true value falling in the set has a controllable lower bound. The chapter uses it as one realization of the calibration idea in machine learning, giving the reader an example of a "prediction that carries its own reliability guarantee."

The methodological roots of judgment, prediction, and the substitution move

27. P. E. Tetlock (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press. [1][2] Tetlock conducts a large-scale, multi-year tracking of expert political forecasting, finding that the long-run accuracy of many experts is unremarkable and often falls short of simple extrapolation baselines. It is the representative work that puts expert judgment under a checkable framework, and gives empirical support to this chapter's "act on a calibrated probability, rather than trust authoritative assertions."

28. P. E. Tetlock & D. Gardner (2015). *Superforecasting: The Art and Science of Prediction.* Crown. [1][4] This book popularizes the research findings of the IARPA forecasting tournament, portraying how the standout "superforecasters" decompose problems, give probabilities, and continually fine-tune with evidence. It leans toward practice, telling exactly how to make judgments that can be checked by calibration in an unverifiable world, well suited for the reader to train their own forecasting habits.

29. G. E. P. Box (1976). "Science and Statistics." *Journal of the American Statistical Association*, 71(356), 791-799. [2][3] This is the source of the line "all models are wrong, but some are useful." Box argues that statistical modeling is an

iterative process of scientific inquiry, which should pursue not absolute correctness but usefulness and improvability. It serves precisely this chapter's contrast of ways of failing: faithful but intractable, or tractable but only approximate.

30. G. Pólya (1945). *How to Solve It: A New Aspect of Mathematical Method*. Princeton University Press. [2][4] Pólya sums up a set of problem-solving heuristics, one of which is "first solve a related, easier problem," then approach the original through it. This is the methodological prototype of the substitution move in this chapter's title, and the reader may see proxy substitution as a generalization of this ancient problem-solving art to settings that cannot be directly verified.

31. H. A. Simon (1956). "Rational Choice and the Structure of the Environment." *Psychological Review*, 63(2), 129-138. [2][4] Simon proposes bounded rationality and satisficing: when capacity and information are limited, an actor does not seek the optimum but stops upon finding a "good enough" option. It provides the theoretical grounding for "replacing the unattainable optimum with a good-enough proxy," and is the root of this chapter's substitution move in decision science.

32. D. Kahneman & S. Frederick (2002). "Representativeness Revisited: Attribute Substitution in Intuitive Judgment." In T. Gilovich, D. Griffin & D. Kahneman (eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*, 49-81. Cambridge University Press. [2] Kahneman and Frederick propose "attribute substitution": when the target attribute is hard to assess, a person unconsciously substitutes a more easily assessed attribute to answer in its place. This is the psychological twin of this chapter's proxy-substitution mechanism, showing that swapping the problem is not only an

engineering strategy but also the default mode in which human intuition operates.

Chapter 12: Contain the Consequences

> **Thesis:** When you cannot prevent the error, manage its consequences. Shrink the damage a wrong, unverified thing can do (decay, before the fact), and make sure that if it does go wrong you will find out (the audit trail, after the fact).

The previous three pairs of moves, even at a lowered standard, were all still trying to get the thing right. This last pair simply concedes that you cannot get it right, and turns instead to managing failure. Since you cannot prevent the error, shrink the damage it can do (decay, before the fact), and make sure that once it happens you can trace it (the audit trail, after the fact).

Decay: Shrink the Blast Radius

The pure form of the first move: give up on guaranteeing that the unverified thing will not fail, and instead fence off, before the fact, how far its failure can reach.

This is the deepest stock-in-trade of computer security. From Saltzer and Schroeder's principle of least privilege in 1975[1], Lampson's confinement in 1973[2], and Dennis and Van Horn's capability mechanism in 1966[7], to Denning's information flow lattice in

1976[5] and the security models of Bell-LaPadula[3] and Biba[4], the theme is the same: grant a component only the minimum capability it needs to do its own job, and nothing more, so that even if it is breached or fails, it cannot raise much of a wave. The sandbox (Goldberg et al., 1996)[9], separation of duties, and defense in depth are all its incarnations. In systems reliability engineering, it is the circuit breaker and the bulkhead (Nygard's *Release It!*[15]), it is blast-radius design, it is canary releases and the error budget (Google SRE[30]); in finance, it is the position limit and the stop-loss; and Taleb's antifragility[17] is likewise about capping the downside.

The unifying idea is this: shift the burden from "make it not fail" (which would require the verification you do not have) to "make it survivable even when it fails." A common quantitative intuition is defense in depth: if k layers of protection each fail independently with probability p, the probability that all of them fail at once is

$$p^k,$$

falling exponentially with the number of layers. But attach at once the warning from Chapter 10: this p^k holds only when the layers fail independently. If the layers fall to the same weakness (the same bypassed kernel, the same administrator password), correlation makes defense in depth degenerate instantly into a single layer. The Fukushima nuclear accident of 2011 is a portrait of just this principle: the plant had multiple redundancy, main power plus backup diesel generators, but a single tsunami drowned them together, and several lines of defense that were supposed to be independent of one another fell to the same cause, leaving defense in depth a sham.

Its standard failure mode lies exactly here: the bypassed decay. Sandboxes have escapes, privileges quietly creep, and what looks

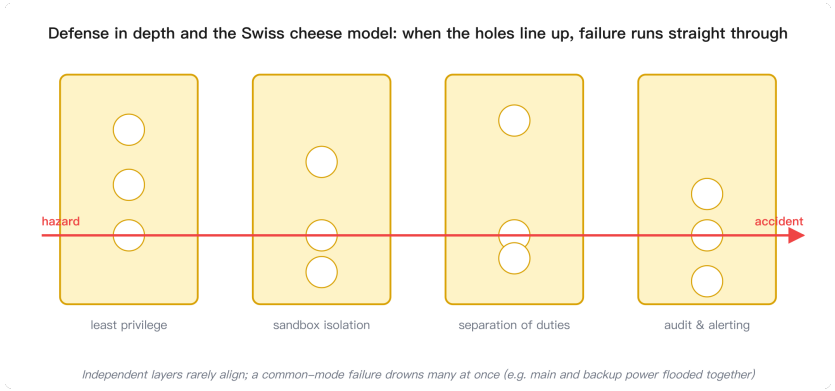

Figure 10: Defense in depth and the Swiss cheese model: when the holes line up, failure runs straight through

like layer upon layer of defense is in fact a set of layers sharing one hidden door. A less often mentioned failure mode is over-protection, which blocks normal function as well, so that people end up working around it to get things done, and security becomes a sham after all.

The Audit Trail: Make the Error Show Itself After the Fact

The pure form of the second move: what you cannot prevent, you make sure to detect the moment it happens. Move the check from before the fact to after it.

Its most hardcore technique comes from cryptography. Merkle's hash tree (Merkle tree) in 1980[18] and Haber and Stornetta's chained timestamps in 1991[19] make a record, once written down, impossible to alter quietly: any change is exposed when the record is checked. Crosby and Wallach's tamper-evident log in 2009[23], Schneier and Kelsey's protection of logs on untrusted machines in 1998[20], and Bellare and Miner's forward-secure signature in 1999[22] make this set of techniques firmer still.

Certificate Transparency (RFC 6962)[24] and Nakamoto's Bitcoin in 2008[27] are, in essence, a global, append-only audit ledger that anyone can verify. Checking whether a record sits within such a tree costs only $O(\log n)$, which is once again the "verifying is cheaper than producing" dividend from Chapter 2.

And this move is in fact far more ancient. Double-entry bookkeeping is one of humanity's earliest tamper-evident ledgers, and Soll, in *The Reckoning*[29], argues that the ability to keep one's own accounts straight bears directly on the rise and fall of nations. Modern financial auditing and independent inspection are the same posture. In science, it is preregistration (Nosek, 2018)[33] and reproducibility (echoing the replication crisis of Chapter 3): registering hypothesis and method before seeing the data, so that the target cannot be moved after the fact.

The unifying idea is this: give up on "stopping the bad thing before the fact" (which takes verification) for "being sure to detect the bad thing after the fact" (which takes only a faithful ledger). Its benefit is twofold: it both lets the error be corrected and, because there is "no getting away," produces a deterrent.

Its standard failure mode is also a single one, yet extremely common: detection that no one responds to. An audit log no one reads, an alert uniformly ignored, is as good as nothing. The Equifax data breach of 2017 is a textbook case: a known vulnerability went unpatched for too long, the intruders lurked in the system for roughly seventy-six days before being noticed, and the personal information of about 147 million people leaked out. The traces were all there in the logs; only, no one looked. Detection without response is a sham. (Another hidden danger is that the log itself can be tampered with, which is precisely what the cryptographic methods above are meant to block.)

All Eight Moves Assembled: The Close of Part III

Take this last pair together: decay shrinks the cost of failure before the fact, the audit trail guarantees failure is found after it. Neither still tries to make the unverified thing correct; instead they reshape the very form of failure itself, one lowering the blast radius, one moving the check to after the fact.

With this, the eight moves are assembled, four pairs complete:

- **Compress the unknown** (Chapter 9): certificate and bound, optimal screening.
- **Borrowed judgment** (Chapter 10): oracle in the loop, redundant consensus.
- **Swap in a problem you can handle** (Chapter 11): proxy substitution, calibration.
- **Contain the consequences** (Chapter 12): the decay fence, the audit trail.

This is that comparison table, the payload of the book. It surfaces again and again across the four sites and in science, dressed in different jargon, yet it is always these eight. By the iron rule laid down in Chapter 4, each move has, as far as possible, accounted for its mechanism, its cross-domain form, and its standard failure mode, rather than resting on surface resemblance alone.

But one sharp question remains unsettled: why these eight, of all things? Is this a list I have cobbled together, or do they each answer to something more basic and unavoidable? If it is only a list, the book is at best a useful manual of classification; if there really is structure behind it, then "convergence" has at last been explained. Part IV chases that question: first attempting to hang the eight moves on a common skeleton, then squarely settling accounts, asking whether this is a law or a very strong empirical pattern.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

1. J. Saltzer & M. Schroeder (1975). *The Protection of Information in Computer Systems.* Proceedings of the IEEE. [2] This survey systematizes the design principles of protection mechanisms, and among them the principle of least privilege becomes the wellspring of the "decay" move: grant a component only the capability it needs to do its own job, and nothing more. The whole line of thought behind this chapter's "shrink the blast radius" originates here, and it is the first required reading for understanding why the scope of failure should be fenced off before the fact.
2. B. Lampson (1973). "A Note on the Confinement Problem." Communications of the ACM. [2] Lampson poses the "confinement problem": how to ensure that a called program cannot leak or misuse the information it has access to, including hard-to-block side channels such as covert channels. It sets a precise problem statement for "caging the thing that goes wrong," which is exactly the core that this chapter's decay move means to address.
3. D. Bell & L. LaPadula (1973). *Secure Computer Systems: Mathematical Foundations.* The MITRE Corporation. [2] The Bell-LaPadula model characterizes confidentiality in a formal way: information may flow only from lower to equal or higher classification, from which comes the famous "no read up, no write down" rule. It demonstrates how the "scope of failure" can be written as a provable lattice structure, a founding work in the theorization of security models.
4. K. Biba (1977). *Integrity Considerations for Secure Com-*

puter Systems. The MITRE Corporation. [2] The Biba model is the dual of Bell-LaPadula, concerned with integrity rather than confidentiality: information may flow only from high trust to low trust, to keep low-trust data from contaminating critical components. Taken together, the two show that the same lattice-theoretic framework can bound the spread of failure from two directions.

5. D. Denning (1976). "A Lattice Model of Secure Information Flow." Communications of the ACM. [2] Denning unifies information-flow security into a single lattice model: data is tagged with security labels, and flow is required to proceed only along the lattice's partial order, thereby constraining where information may go statically, at compile time or run time. It supplies a common mathematical language for the preceding security models, the theoretical core of information flow control.
6. D. Clark & D. Wilson (1987). "A Comparison of Commercial and Military Computer Security Policies." IEEE Symposium on Security and Privacy. [2] Clark and Wilson point out that commercial settings care more about integrity than about military-style confidentiality, and propose an integrity model centered on well-formed transactions and separation of duties. It extends "decay" from military classification levels to everyday settings such as commercial accounting, showing that the form of bounding the scope of failure varies with the domain.
7. J. Dennis & E. Van Horn (1966). "Programming Semantics for Multiprogrammed Computations." Communications of the ACM. [2] This early paper introduces the concept of the capability: access rights attach directly to a reference in the form of an unforgeable token, and only holding the token lets one operate on the object. It is the wellspring of the capability security model, providing a mechanism-level implementation path for "granting only the minimum

capability needed."

8. N. Provos, M. Friedl & P. Honeyman (2003). "Preventing Privilege Escalation." 12th USENIX Security Symposium. [2] The authors discuss how to use privilege separation to isolate privileged operations into a minimal, trusted slice of code, so that the main program, even if breached, can act only at low privilege; OpenSSH's privilege separation is the representative practice. It brings least privilege down to the engineering detail of real systems.
9. I. Goldberg, D. Wagner, R. Thomas & E. Brewer (1996). "A Secure Environment for Untrusted Helper Applications." 6th USENIX Security Symposium. [2] This paper introduces Janus, which uses system-call interception to build a restricted runtime environment for untrusted programs, an early exemplar of the user-space sandbox. This chapter lists the sandbox as an incarnation of the decay move, and this paper is exactly the representative source of the sandbox idea.
10. C. Perrow (1984). *Normal Accidents: Living with High-Risk Technologies*. Basic Books. [2][1] Perrow advances the "normal accidents" thesis: in highly complex, tightly coupled systems, catastrophic accidents are not accidental but structurally inevitable, and cannot be rooted out by adding protections. It supports this chapter's stance from the other side: when errors cannot be prevented one by one, the center of gravity should move to managing consequences rather than vainly trying to abolish failure.
11. N. Leveson (2011). *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press. [2] Leveson reconstructs safety engineering with systems theory, proposing the STAMP model, which treats an accident as a failure of the control structure rather than the fault of a single part, and stresses using constraints to bound hazardous states. It supplies a systems-level methodology for "fencing off the

scope of failure before the fact."

12. J. Reason (1990). *Human Error.* Cambridge University Press. [2] Reason analyzes human error systematically and proposes the famous "Swiss cheese model": every layer of protection has holes, and only when the holes in multiple layers happen to line up does an accident run straight through. It is the very source of this chapter's defense-in-depth intuition, and also a reminder that once the holes in the layers become correlated, the multiple layers degenerate into one.
13. E. Hollnagel, D. Woods & N. Leveson (2006). *Resilience Engineering: Concepts and Precepts.* Ashgate. [2][4] This collection lays the foundations of "resilience engineering": a system's safety lies not in eliminating faults but in having the capacity to absorb disturbances and to keep running and recover after failure. It accords closely with this chapter's theme, shifting the center of gravity explicitly from "not failing" to "surviving even when it fails."
14. A. Avizienis, J.-C. Laprie, B. Randell & C. Landwehr (2004). "Basic Concepts and Taxonomy of Dependable and Secure Computing." IEEE Transactions on Dependable and Secure Computing. [2] This widely cited taxonomy clarifies the chain of fault, error, and failure, and the relations among means such as fault tolerance, fault prevention, and fault detection. It provides an agreed terminological framework for the various decay and audit-trail methods this chapter discusses, suitable as a reference for conceptual calibration.
15. M. Nygard (2007). *Release It! Design and Deploy Production-Ready Software.* Pragmatic Bookshelf. [2][4] Nygard writes reliability engineering as a practical handbook, proposing stability patterns such as the circuit breaker, the bulkhead, timeouts, and compartment isolation, to keep a local fault from cascading into a global collapse. The main text takes it as the representative of blast-radius design, a direct read on applying the decay

idea to production systems.

16. R. Anderson (2020). *Security Engineering: A Guide to Building Dependable Distributed Systems* (third edition). Wiley. [2][4] Anderson's monumental work ranges across cryptography, access control, economic incentives, and real-world offense and defense, an authoritative comprehensive textbook of security engineering. Almost every topic this chapter touches, least privilege, auditing, tamper-evidence, can be found there in fuller development, suitable as a base text to read through.
17. N. N. Taleb (2012). *Antifragile: Things That Gain from Disorder*. Random House. [4] Taleb introduces "antifragility": some systems not only survive volatility but benefit from it, the key being to cap the downside and preserve the upside. This chapter cites it to show that the extreme posture of decay is to put a ceiling on loss, a perspective for understanding "contain the consequences" from the angle of risk management.
18. R. Merkle (1980). "Protocols for Public Key Cryptosystems." IEEE Symposium on Security and Privacy. [2] Merkle here proposes the idea of tree-structured authentication using a hash tree (the Merkle tree): a large body of data is merged into a single root hash, and the authenticity of any one item can be checked with only an $O(\log n)$ path. It is the technical bedrock of this chapter's "audit trail" move, and the common ancestor of the later Certificate Transparency and blockchain.
19. S. Haber & W. S. Stornetta (1991). "How to Time-Stamp a Digital Document." Journal of Cryptology. [2] The two authors propose chaining document timestamps together with hashes, so that any later tampering breaks the continuity of the chain and is exposed. This is the pioneering work on chained tamper-evident records, directly inspiring the later blockchain structure, and the key to understanding why the

audit trail "cannot be altered."

20. B. Schneier & J. Kelsey (1998). "Cryptographic Support for Secure Logs on Untrusted Machines." 7th USENIX Security Symposium. [2] This paper designs a scheme for protecting logs on machines that may be compromised: even if an attacker later gains control, they cannot delete or alter the earlier records without being noticed. It advances tamper-evident logging into untrusted environments, one of the hardcore techniques of this chapter's audit-trail move.
21. B. Schneier & J. Kelsey (1999). "Secure Audit Logs to Support Computer Forensics." ACM Transactions on Information and System Security. [2] This is the journal-version extension of the previous work, treating more fully the construction of secure audit logs that support forensics. It shows that the audit trail must not only record faithfully but also withstand adversarial checking after the fact, matching this chapter's goal of "making the error show itself after the fact."
22. M. Bellare & S. Miner (1999). "A Forward-Secure Digital Signature Scheme." CRYPTO '99. [2] Bellare and Miner propose the forward-secure signature: the key evolves periodically, so that even if the current key is leaked, an attacker cannot forge signatures from earlier periods. It provides a key safeguard for the audit trail, so that past records remain unimpersonable and untamperable even after the private key is compromised.
23. S. Crosby & D. Wallach (2009). "Efficient Data Structures for Tamper-Evident Logging." 18th USENIX Security Symposium. [2] The authors design a tamper-evident log structure that can be appended to and audited efficiently, letting a verifier confirm the integrity and consistency of records without trusting the log server. It integrates the foregoing cryptographic methods into a deployable data structure, a representative work in the engineering of the audit-trail

move.

24. B. Laurie, A. Langley & E. Kasper (2013). *RFC 6962: Certificate Transparency.* IETF. [2] Certificate Transparency records all issued TLS certificates in a public, append-only Merkle log auditable by anyone, leaving mis-issued or malicious certificates nowhere to hide. It is the real-world model for this chapter's "global audit ledger that anyone can verify," showing how the audit trail can be deployed at scale.
25. L. Lamport, R. Shostak & M. Pease (1982). "The Byzantine Generals Problem." ACM Transactions on Programming Languages and Systems. [2] This classic paper formalizes the Byzantine fault tolerance problem: when some nodes may behave arbitrarily badly, how can the honest nodes agree on a value, and it gives the theoretical bound on how many faulty nodes can be tolerated. It is the consensus-theoretic foundation on which the publicly verifiable ledger rests, listed by this chapter under "theoretically studied material."
26. M. Castro & B. Liskov (1999). "Practical Byzantine Fault Tolerance." 3rd USENIX Symposium on Operating Systems Design and Implementation (OSDI). [2] Castro and Liskov give the first Byzantine fault tolerance algorithm, PBFT, practical in a real asynchronous network, bringing theoretical consensus to engineerable performance. It shows that a ledger anyone can verify and that tolerates malicious nodes is no fantasy, providing implementation support for the trusted basis of the audit trail.
27. S. Nakamoto (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System.* White paper. [2] Nakamoto's white paper proposes Bitcoin: a decentralized, append-only blockchain driven by proof of work, letting mutually distrusting parties agree on the transaction history. This chapter views it in essence as a globally verifiable audit ledger, the extreme realization of the audit-trail idea on an open network.

28. D. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler & G. Sussman (2008). "Information Accountability." Communications of the ACM. [2][4] The authors argue for shifting from "blocking access before the fact" to "accountability after the fact": allow information to flow, but require that its uses be auditable and that violations be traceable and answerable. This is wholly isomorphic to this chapter's "audit-trail" posture of moving the check from before to after the fact, a programmatic statement of the idea in privacy governance.
29. J. Soll (2014). *The Reckoning: Financial Accountability and the Rise and Fall of Nations.* Basic Books. [1] Soll argues from financial history that whether a regime can keep and faithfully present its own accounts bears directly on its rise and fall, with double-entry bookkeeping the key accountability technique among them. It traces the lineage of the audit trail back to humanity's earliest tamper-evident ledgers, showing that the power of the audit ledger is of long standing.
30. B. Beyer, C. Jones, J. Petoff & N. Murphy (2016). *Site Reliability Engineering: How Google Runs Production Systems.* O'Reilly. [4] This book introduces Google's SRE practice systematically, including the error budget, canary releases, monitoring and alerting, and controlled failure drills. This chapter borrows its error budget and canary releases to show how decay can be institutionalized in large-scale production, a modern model for the engineering of "contain the consequences."
31. J. Ioannidis (2005). "Why Most Published Research Findings Are False." PLoS Medicine. [3] Ioannidis argues with statistical modeling that under conditions of low priors, small samples, multiple comparisons, and excessive researcher degrees of freedom, a great many published findings are likely false positives. This is an analytical

argument rather than an empirical replication study, supplying a problem diagnosis for the preregistration and reproducibility this chapter mentions.

32. Open Science Collaboration (2015). "Estimating the Reproducibility of Psychological Science." Science. [3] This is a large-scale empirical effort: many groups of researchers attempt to replicate over a hundred psychology studies, and a sizable share fail to reproduce. It turns Ioannidis's theoretical worry into visible data, the landmark evidence for the "replication crisis," echoing this chapter's emphasis on after-the-fact verifiability.
33. B. Nosek, C. Ebersole, A. DeHaven & D. Mellor (2018). "The Preregistration Revolution." PNAS. [3] Nosek and colleagues advocate preregistration: publicly registering hypothesis and analysis method before seeing the data, separating exploratory from confirmatory research so the target cannot be moved after the fact. It is the "audit trail" practice in science, shifting verification from trust before the fact to checking after it, corresponding directly to this chapter's theme.

Part IV: Levers

Chapter 13: The Eight Levers

> **Thesis:** The eight moves are not an arbitrary list; each one pulls a different lever within the decomposition of risk and information, and that is exactly why they feel "complete."

Part III handed over that table: eight moves, four pairs, appearing again and again under different jargon across four concrete sites plus science. But a list, however tidy, is still only a list. What this chapter presses on is this: why these eight, and not others? Did I assemble them, or does each of them lodge at some unavoidable position? If the latter, then the convergence has been explained; otherwise this book is at best a handy manual of classification.

I want to put forward a candidate explanation. Let me say the unflattering part first: it is an organizing scheme, not a proof. When you have finished the chapter, please bring along that skeptical knife from Chapter 14.

A Crude Decomposition

Strip "acting under unverifiability" down to its barest form, and what you are really managing is risk. Borrowing the old language of decision theory (Wald[1], Savage[4], von Neumann and

Morgenstern[3]), risk can be written, roughly, as

$$\text{Risk} \;\approx\; \Pr(\text{fail}) \;\times\; \text{Cost}(\text{fail}),$$

and all of this proceeds under an information budget B: the checks, samples, computation, and time you can spend to cut down uncertainty are all finite.

The formula looks simple, but the point is this: the places on its right-hand side where you can intervene are a countable few. You can alter the definition of "failure" itself, or the probability of failure, or your knowledge of that probability, or the cost of failure, or how this information budget is spent, or the timing at which checking happens. My proposition is: the eight moves occupy exactly one position each, with no ninth slot left to fill.

Eight Moves, Eight Positions

Match each move to the lever it pulls:

Move	The lever it pulls	Its position in the decomposition of risk
proxy substitution	changes the target you measure and optimize	rewrites the definition of "failure" itself
certificate / bound	presses uncertainty into a guaranteed bound on one slice	drives Pr(fail) near zero locally
oracle in the loop	brings in a verifying power you do not have on your own	lowers Pr(fail) with outside help

Move	The lever it pulls	Its position in the decomposition of risk
redundancy / consensus	makes the failures of several judgments decorrelate	lowers the joint failure probability Pr(all fail)
optimal screening	spends the information budget where the marginal return is highest	allocates B to maximize the cut in uncertainty
calibration	puts a truthful price on residual risk	makes Pr(fail) known, so you can bet on it
decay / fencing / containment	shrinks the blast radius	lowers Cost(fail)
audit trail	moves checking from before the fact to after it	shifts the timing of checking, turning an irrecoverable failure into a recoverable one

Read this table, and the feeling that it is "an assembled list" should loosen a little. The eight moves are not eight tools gathered at random; they take up, between them, the positions of "Pr, knowledge of Pr, cost, budget allocation, timing of checking, definition of the target," filling almost one by one the places where that decomposition can be acted upon. The proposition can then be put this way: if these really are all the levers there are, then this set of moves is complete, and the convergence is thereby explained, since any capable actor will sooner or later rediscover them, because there is nothing else to pull.

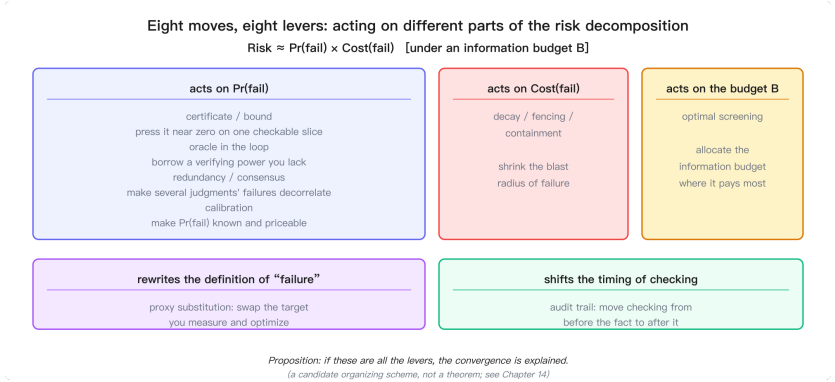

Figure 11: The eight moves mapped onto different parts of the decomposition of risk (a candidate organizing scheme, not a theorem)

Why This Can Cross Substrates

If the above holds, it explains, in passing, the puzzle the book opened with: why mathematicians, engineers, and organizations would, without prior agreement, reach for these same eight moves.

In studying vision, Marr[17] distinguished three levels: the computational level (what problem is to be solved, under what constraints), the algorithmic level (what representations and processes are used), and the implementational level (what hardware it runs on). The eight moves live at the computational level; they are the answer to "given the constraint of unverifiability, which few places can logically still be acted upon," and that answer does not depend on whether you are a carbon-based mathematician, a silicon-based program, or a bureaucracy made of people. The substrate varies wildly, yet the constraint at the computational level is one and the same, so the responses converge. Simon's bounded rationality[12] and his "sciences of the artificial"[13] speak of precisely this kind of behavior: shaped by the constraints of the environment rather than by the internal makeup of the actor.

Here we must also bring in the no free lunch theorem (Wolpert and Macready[25]). It says: averaged over all possible problems, no method outperforms another. This knife cuts both ways. On one side, it supports the book's restraint: there is no universal solution, you must lean on the specific structure of the problem to choose a lever, which is exactly why the five faces must be treated separately. On the other side, it warns: anyone who claims to have "found the unifying key," myself included, should rein in some of their pride. When you cannot even pin down the probability of failure, the levers grow robust versions: Gilboa and Schmeidler's maxmin expected utility[20], Hansen and Sargent's robust control[30], decision-making under Knightian uncertainty, are all moves that still mount a defense against the worst case even when Pr itself is ambiguous.

A Strong Claim That Must Be Amplified

Now amplify the unflattering part.

The scheme above is a candidate organizing structure, not a theorem. That decomposition of risk is informal; I have given no rigorous model of actor and environment, and so I cannot prove that the optimal strategy is exactly these eight levers. "These are all the levers there are" is an assertion, not an established result. I have no evidence that this table is exhaustive, nor can I rule out that it is merely a post hoc framework: a narrative flexible enough to stuff many sets of moves into. Some of the assignments in the table (for instance, redundancy both lowers joint failure and looks like a special kind of screening) even overlap, which itself shows that this decomposition is not yet clean.

I place it here because it has organizing force and explanatory appeal, not because it has been proved. It meets the standard of a good conjecture: clear, refutable, able to unify a large mass of phenomena. But it has not yet risen to a theorem.

So the final question becomes unavoidable: is this cross-domain convergence a law forced out by something, or merely a strong but, in the end, empirical pattern? The next chapter settles the account with it, head-on and squarely.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

1. A. Wald (1950). *Statistical Decision Functions.* John Wiley & Sons. [2][4] Wald recast statistical inference as a decision problem against nature: the decision-maker must choose a strategy under risk (the expectation of loss), using the minimax criterion of minimizing the maximum risk to cope with the unknown state. This book founded statistical decision theory, and the scholarly root of this chapter's crude decomposition, "risk is roughly failure probability times cost," lies right here.
2. A. Wald (1939). "Contributions to the Theory of Statistical Estimation and Testing Hypotheses." *The Annals of Mathematical Statistics*, 10(4), 299-326. [2] This is Wald's early paper unifying estimation and testing within a loss-function framework, preceding his later monograph, already introducing the ideas of a risk function and a least favorable prior. It marks the starting point of the turn toward "talking about statistics in the language of decision," and is valuable for understanding why this chapter draws on decision theory.
3. J. von Neumann and O. Morgenstern (1944). *Theory of Games and Economic Behavior.* Princeton University Press. [2] Von Neumann and Morgenstern founded game theory

and derived the expected utility theorem from a set of axioms: preferences satisfying the rationality axioms can be represented as maximizing the expectation of a utility. This is the normative baseline for "betting under uncertainty," cited here as one source of the chapter's calculus of risk.

4. L. J. Savage (1954). *The Foundations of Statistics*. John Wiley & Sons. [2][4] Savage, using a set of axioms about preferences over actions, derived subjective probability and utility together, building Bayesian decision theory on personalist probability. It is the founding work of the modern framework of "subjective probability plus expected utility," and also the target that Ellsberg's paradox later challenges.
5. F. H. Knight (1921). *Risk, Uncertainty and Profit*. Houghton Mifflin. [2] Knight distinguished quantifiable "risk" from "uncertainty" to which no probability can be assigned, and argued that entrepreneurial profit comes precisely from bearing the latter. This distinction is the key conceptual source when this chapter speaks of "cannot even pin down the probability of failure," and Knightian uncertainty runs through the later series of robust decision works.
6. J. M. Keynes (1921). *A Treatise on Probability*. Macmillan. [2] Keynes developed a logical interpretation of probability, treating it as a rational degree of belief between propositions, and stressed that many probabilities are neither numerical nor necessarily comparable. It laid the groundwork for later non-additive, imprecise probabilities, reminding the reader that the epistemic status of probability itself is far more complex than any formula.
7. F. P. Ramsey (1931). "Truth and Probability." *The Foundations of Mathematics and other Logical Essays* (R. B. Braithwaite, ed.). Kegan Paul, Trench, Trubner & Co., 156-198. [2] Ramsey was the first to argue that a person's degree of belief can be measured operationally through their bet-

ting behavior, and that avoiding a sure-loss combination (a Dutch book) requires those degrees of belief to obey the axioms of probability. This is the founding work of subjective probability, and it provides the philosophical and operational basis for this chapter's "putting an honest price on residual risk."

8. B. de Finetti (1937). "La prévision: ses lois logiques, ses sources subjectives." *Annales de l'Institut Henri Poincaré*, 7(1), 1-68. [2] De Finetti proposed subjective probability and backed it with the Dutch-book argument and the representation theorem for exchangeability, arguing that probability is "only" a coherent personal degree of belief. Together with Ramsey it forms the cornerstone of Bayesianism, an essential source for understanding calibration and honest pricing.
9. F. J. Anscombe and R. J. Aumann (1963). "A Definition of Subjective Probability." *The Annals of Mathematical Statistics*, 34(1), 199-205. [2] The two authors, by introducing objective randomizing devices (such as roulette lotteries), gave an axiomatization of subjective probability and utility more compact than Savage's. This framework later became the standard stage for ambiguity decision theory, and the several ambiguity-aversion works cited in this chapter are built upon it.
10. D. Ellsberg (1961). "Risk, Ambiguity, and the Savage Axioms." *The Quarterly Journal of Economics*, 75(4), 643-669. [2] Ellsberg used two famous ball-drawing experiments to show that people systematically prefer known probabilities and avoid "ambiguity," a behavior that violates Savage's axioms and cannot be explained by any single subjective probability. It established ambiguity as an independent phenomenon, and is the direct motivation for the subsequent robust and multiple-prior theories.
11. R. D. Luce and H. Raiffa (1957). *Games and Decisions:*

Introduction and Critical Survey. John Wiley & Sons. [2][4] This book is a classic introduction to, and critical survey of, game theory and decision theory, both laying out expected utility and game solution concepts clearly and candidly discussing the limits of applicability of each axiom. It suits the reader as a general entry point into the chapter's decision-theoretic background, at once systematic and critical.

12. H. A. Simon (1955). "A Behavioral Model of Rational Choice." *The Quarterly Journal of Economics*, 69(1), 99-118. [2] Simon here proposed bounded rationality and "satisficing": an actor limited by cognition and information does not seek the global optimum, but searches until it finds an option good enough and then stops. This is the core support for this chapter's argument that "constraints at the computational level shape behavior," explaining why different substrates converge on the same set of responses.
13. H. A. Simon (1969). *The Sciences of the Artificial*. MIT Press. [2][3] Simon argued that the behavior of an artifact is determined more by the constraints of the environment it inhabits than by its internal construction, and called for founding "design" as a discipline of artificial systems. This chapter draws on it to argue that the eight moves live at Marr's computational level, being shaped by the constraints of the environment rather than the internal makeup of the actor.
14. K. R. Popper (1959). *The Logic of Scientific Discovery*. Hutchinson. [2][3] Popper systematically proposed falsificationism: a scientific theory cannot be empirically verified, only refuted, so falsifiability becomes the line between science and non-science. When this chapter admits at the close that its decomposition "rises to a good conjecture but not yet to a theorem," it is using precisely this measure of refutability.
15. A. Tversky and D. Kahneman (1974). "Judgment under Un-

certainty: Heuristics and Biases." *Science*, 185(4157), 1124-1131. [2] Tversky and Kahneman documented the heuristics people use in judging probability (representativeness, availability, anchoring) and the systematic biases they bring. It shows how real actors deviate from the Bayesian ideal, forming a contrast with moves such as calibration and optimal screening that require an honest estimate of probability.

16. D. Kahneman and A. Tversky (1979). "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47(2), 263-291. [2] Prospect theory holds that people evaluate outcomes by gains and losses relative to a reference point, rather than by final wealth, are more sensitive to losses, and distort probability weights in a nonlinear way. It is a descriptive correction to expected utility, reminding the reader that cost and probability do not multiply neutrally in real decisions.
17. D. Marr (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman. [2][3] Marr proposed three levels for analyzing an information-processing system: the computational level (what problem is to be solved), the algorithmic level (what representations and processes are used), and the implementational level (what hardware it runs on). This chapter draws on exactly this hierarchy to argue that the eight moves live at the computational level and can therefore cross carbon-based, silicon-based, and organizational substrates.
18. J. O. Berger (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag. [2][4] Berger systematically organized the statistical decision theory at the meeting point of the Bayesian and frequentist schools, covering loss functions, risk, admissibility, and robust Bayesian analysis. It is the standard reference for grounding this chapter's informal decomposition of risk in rigorous statistical language.

19. D. E. Bell, H. Raiffa and A. Tversky (eds.) (1988). *Decision Making: Descriptive, Normative, and Prescriptive Interactions*. Cambridge University Press. [2][4] This collection is organized around three orientations in decision research: descriptive (how people actually decide), normative (how rationality ought to proceed), and prescriptive (how to help people decide better), and discusses how the three interact. It gives the reader a map for placing the work of each school, echoing this chapter's repeated weighing between the normative and the empirical.
20. I. Gilboa and D. Schmeidler (1989). "Maxmin Expected Utility with Non-Unique Prior." *Journal of Mathematical Economics*, 18(2), 141-153. [2][4] Gilboa and Schmeidler gave an axiomatization for ambiguity decision-making: the actor holds a set of priors and evaluates an action by the least favorable among them, that is, a maxmin expected utility over the set of priors. This is exactly the move this chapter describes as "still mounting a defense against the worst case even when Pr itself is ambiguous," a representative formalization of robust decision-making.
21. D. Schmeidler (1989). "Subjective Probability and Expected Utility without Additivity." *Econometrica*, 57(3), 571-587. [2] Schmeidler introduced non-additive subjective probability (capacities) and the corresponding Choquet expected utility, making ambiguity aversion representable in a consistent way. Together with the preceding entry it is a founding work of ambiguity decision theory, loosening, along another technical route, the constraint that probability must be additive.
22. P. Walley (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall. [2][4] Walley systematically developed the theory of imprecise probabilities, characterizing belief under insufficient evidence with upper and lower probabilities (or a set of probabilities), and gave the corre-

sponding criteria for coherence and inference. It provides a complete statistical language for the predicament of "cannot even pin down the probability," and is the deep theoretical backing for the calibration and robustness lines of thought.

23. G. Gigerenzer and D. G. Goldstein (1996). "Reasoning the Fast and Frugal Way: Models of Bounded Rationality." *Psychological Review*, 103(4), 650-669. [2] Gigerenzer and Goldstein argued that simple "fast and frugal" heuristics can, in real environments, often match or even outperform complex models, exhibiting a kind of ecological rationality. It completes Simon's bounded rationality from the positive side, explaining why simple rules may be exactly the reasonable choice when the information budget is limited.
24. D. H. Wolpert (1996). "The Lack of A Priori Distinctions between Learning Algorithms." *Neural Computation*, 8(7), 1341-1390. [2] Wolpert proved the no free lunch result for supervised learning: averaged over all possible target functions, any learning algorithm has the same generalization performance. It shows that there is no universal learner independent of problem structure, and is the source, on the learning side, of this chapter's no free lunch argument.
25. D. H. Wolpert and W. G. Macready (1997). "No Free Lunch Theorems for Optimization." *IEEE Transactions on Evolutionary Computation*, 1(1), 67-82. [2] Wolpert and Macready extended no free lunch to optimization: averaged over all possible objectives, no optimization algorithm outperforms another. This chapter uses this double-edged knife both to support the restraint that "one must lean on problem structure to choose a lever" and to warn anyone claiming to have found the unifying key to rein in their pride.
26. I. Gilboa and D. Schmeidler (2001). *A Theory of Case-Based Decisions*. Cambridge University Press. [2][4] The two authors proposed case-based decision theory: when an actor cannot articulate the state space and probability is

out of the question, decisions can be driven by recall of and analogy to past similar cases. It provides another normative model for the predicament in which the probability framework breaks down entirely, broadening this chapter's imagination of how deep "unverifiable" can go.

27. T. F. Bewley (2002). "Knightian Decision Theory. Part I." *Decisions in Economics and Finance*, 25(2), 79-110. [2][4] Bewley formalized Knightian uncertainty as the incompleteness of preferences: when the evidence is insufficient to rank two options, the actor may decline to choose, supplemented by an inertia assumption that maintains the status quo. It gives a clean axiomatic expression to "uncertainty that cannot be priced," and is the modern continuation of this chapter's Knightian theme.
28. P. Klibanoff, M. Marinacci and S. Mukerji (2005). "A Smooth Model of Decision Making under Ambiguity." *Econometrica*, 73(6), 1849-1892. [2] The three authors proposed a "smooth" model of ambiguity decision-making: by layering a further utility function over a second-order distribution on priors, it separates ambiguity attitude from risk attitude and avoids the non-smooth kink of maxmin. It makes the degree of ambiguity aversion tunable and analyzable, a finer notch within the spectrum of robust decision-making.
29. F. Maccheroni, M. Marinacci and A. Rustichini (2006). "Ambiguity Aversion, Robustness, and the Variational Representation of Preferences." *Econometrica*, 74(6), 1447-1498. [2][4] The authors gave the unified representation of variational preferences, gathering maxmin expected utility, multiplier (robust control) preferences, and others as special cases, with ambiguity attitude characterized by a penalty term on deviation. It mathematically links the several robust moves mentioned in this chapter into one family, its value lying in revealing their common skeleton.

30. L. P. Hansen and T. J. Sargent (2008). *Robustness*. Princeton University Press. [2][4] Hansen and Sargent brought robust control from control theory into economic decision-making: the decision-maker distrusts the model in hand and so optimizes against the least favorable among a family of nearby models, seeking robustness to specification error. This is the main source of this chapter's "robust control" move, displaying a dynamic form of mounting a defense against the worst case.
31. P. P. Wakker (2010). *Prospect Theory: For Risk and Ambiguity*. Cambridge University Press. [2] Wakker systematized and axiomatized prospect theory, handling decision weights under risk and ambiguity in a unified way, and provided operational methods of measurement. It stitches descriptive prospect theory together with normative ambiguity theory, and is the authoritative monograph for the reader going deeper into the theme of probability weighting.
32. I. Gilboa and M. Marinacci (2013). "Ambiguity and the Bayesian Paradigm." *Advances in Economics and Econometrics: Theory and Applications, Tenth World Congress* (D. Acemoglu, M. Arellano and E. Dekel, eds.). Cambridge University Press. [2][4] This survey traces the whole line of ambiguity decision-making and faces a fundamental question head-on: in what sense the Bayesian paradigm suffices, and where it needs to be replaced by models such as multiple priors. It is the best navigation into this chapter's cluster of ambiguity and robustness literature, and matches this chapter's restraint in attitude.
33. P. Bossaerts and C. Murawski (2017). "Computational Complexity and Human Decision-Making." *Trends in Cognitive Sciences*, 21(12), 917-929. [2] The two authors argued that many real decision problems are intractable in computational complexity (such as the knapsack and other NP-hard problems), and that the brain's performance and

strategies are shaped by this hard constraint. It provides hard evidence for bounded rationality from the angle of computational complexity, echoing this chapter's core claim that "constraints at the computational level force convergence."

Chapter 14: Theorem or Pattern?

> **Thesis:** The book's central reckoning. Is this cross-domain convergence a theorem (something forces any bounded actor facing the unverifiable to walk into these moves), or merely a strong empirical pattern (we keep seeing it, yet have not proved it must be so, and a selection effect may explain the rhyme)?

This is the book's chapter of reckoning. It puts the question I have carried all along squarely on the table:

Is this cross-domain convergence a law (something forces any bounded actor facing the unverifiable to walk, of necessity, into these moves), or merely a very strong empirical pattern (we keep seeing it, yet have not managed to prove it cannot be otherwise, and a selection effect may by itself suffice to explain the rhyme)?

I will try to make both sides as hard as I can, including the side that cuts against me. If this chapter leans at all, it should lean toward doubt.

The Side for "Law"

The first thread is that lever decomposition from the previous chapter. If the eight moves really do occupy every position the risk decomposition leaves open to attack, then the convergence is not coincidence but something forced out by structure: any capable actor will sooner or later rediscover them, because there is no other choice. If this argument holds, its weight is great.

The second thread is independent rediscovery. In the sociology of science, Merton studied the phenomenon of "multiple discovery": the same idea is often arrived at, almost simultaneously and independently, by people who knew nothing of one another[12]. The calculus had Newton and Leibniz; natural selection had Darwin and Wallace; the patent applications for the telephone saw Bell and Gray file on the very same day. Examples like these crowd the history of science in numbers too large to look like chance. The eight moves of this book were likewise reinvented again and again across cryptography, statistics, number theory, organization theory, and security engineering, fields that at the time were not much in communication. If a thing keeps being stumbled upon independently, the scent is more of necessity than of borrowing.

The third thread comes from a firm precedent within science. The renormalization group and universality classes in physics show that systems of wildly different structure converge, near a critical point, to exactly the same behavior, and there it really is a theorem[17,18]. One astonishingly concrete fact: the behavior of a liquid near its critical point and the behavior of a magnet near its Curie point are described by the same set of "critical exponents." Molecules and magnetic moments have nothing to do with each other, yet they fall into the same universality class, because what governs critical behavior is coarse-grained features like symmetry and dimension, not microscopic detail. Wimsatt's robustness analysis says a conclusion that can be reached again and again

by many mutually independent routes is more likely to be true[20]. Whewell's 1840 "consilience of inductions"[11] and Wigner's essay "The Unreasonable Effectiveness of Mathematics" [22] both speak to the credibility that such cross-domain convergence brings. Convergence has always been one of science's ancient signals that "this is the real thing."

The Side for "Pattern, or Weaker"

Now for what cuts against me, and I think this side weighs no less than the other.

The most lethal blow is that those fields are not nearly as independent as they look. They share one mathematical substrate, probability, optimization, information theory, soaking into the roots of every field; they share one human cognition, since these disciplines were all built by the same kind of brain; and they borrow and cite from one another, never sealed off, with Shannon's information theory flowing into almost everything and the language of decision theory diffusing everywhere. If the convergence is only because everyone draws from the same mathematical toolbox, is shaped by the same kind of mind, and has been imitating one another all along, then "independent rediscovery" is heavily discounted, and the so-called convergence may be only the echo of a common origin.

Second, that lever decomposition may well be an after-the-fact frame. Is it perhaps merely flexible enough to fit many sets of moves inside it? The table in the previous chapter already showed its hand: redundancy was filed under "decorrelating joint failure" yet also looked like a special case of screening, and the overlap in placement shows that the ruler itself is not hard enough. Dennett raised a key question: when does a pattern count as real, rather than imposed[13]? The criterion is its predictive and compressive power; a real pattern lets you predict something new. My

framework, so far, has mainly organized known moves rather than predicted a move never before seen and later confirmed. That is a test it has not yet passed.

Third, the selection effect. I went into these fields with eyes set on "finding convergence," so I may well have unconsciously filtered out the fields and counterexamples that did not fit, keeping only what rhymed. The replication crisis is the loudest alarm of all: an empirical pattern that looks utterly robust may be only the product of systematic bias[29,30]. I have no reason to assume I am immune to such bias.

Fourth, even if the convergence is real, it need not point to a deep law. Laudan's pessimistic meta-induction reminds us that one successful theory after another in history was later overturned[6]; converging on a pattern is not the same as converging on truth. Cartwright argues that the fundamental laws do not in fact describe the world truthfully[14], and Anderson's "more is different" points out that higher levels have laws of their own[15]; perhaps this book's convergence is no more than a regularity at the scale of a "special science," not some fundamental law. Worrall's structural realism offers a middle path[7]: perhaps what truly survives is the structure (these levers themselves), even if the gloss I wrap around it is wrong.

What It Would Take to Settle It

To truly close this question, one would need something I cannot supply: a formal model of "a bounded actor plus an unverifiable system," together with a theorem proving that, under that model, the optimal, or the unique, strategy is exactly these few levers. To my knowledge, no such model yet exists. The closest real theorem is no free lunch[24], and it points, of all directions, the other way: there is no universally dominant method. Until that model and that theorem appear, the question is open, and I do not intend to

pretend it has closed.

So let me state plainly what this book delivers: it is a conjecture, a strong, useful conjecture whose boundaries have been drawn plainly, together with a shared vocabulary that can string many fields together. It is not a theorem. This is precisely the posture the book has recommended throughout, a calibrated belief, not a binary verdict.

A Recursive Close

Here, a thing long buried finally surfaces: a book about how to act under unverifiability cannot verify its own central claim.

This sounds like an awkward self-reference, but it is in fact the book's most candid moment. Facing a central claim it cannot itself verify, it has no other choice but to do the very thing it describes throughout. It states a calibrated belief (I think this convergence is real, but cannot give a proof). It draws the boundaries of the claim (this is a conjecture, not a theorem). It openly invites refutation (go find a field that does not converge, go find the counterexample where one move breaks the decomposition; that would be its black swan). And then it goes on speaking anyway, goes on handing you this vocabulary, because it is useful, even if unproven.

In other words, the book has personally rehearsed its own set of moves: it used proxy substitution (replacing "prove the convergence" with "exhibit and organize the convergence"), it used calibration (pricing its own confidence), and it used a falsificationist audit trail (writing down, in black and white, an assertion that can be overturned). A self-referential system cannot fully justify itself from within, a fate we ought long since to have grown used to[26,27,25]. But the inability to prove itself from within does not mean it cannot act. If the book's thesis is right, then this "writing

itself in the very way it describes" is not a defect but a faint yet fitting piece of corroborating evidence.

So a last question follows. If verification is usually unavailable, and even this book can offer only an unproven belief, then what exactly is that great heap of things we ordinarily call "knowledge"? The next chapter sets its landing point on epistemology.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

1. T. S. Kuhn (1962). *The Structure of Scientific Revolutions*. University of Chicago Press. [3][1] Kuhn argues that science does not advance by linear accumulation but alternates between normal science (solving puzzles within an existing paradigm) and scientific revolution (the replacement of paradigms), with "incommensurability" lying between old and new paradigms. The first edition appeared in 1962, also published as a separate volume of the *International Encyclopedia of Unified Science*. It is a founding text for understanding "how science progresses," and it prompts this chapter's very question: is cross-domain convergence shaped by a single paradigm, or forced out independently?

2. K. R. Popper (1959). *The Logic of Scientific Discovery*. Hutchinson. [3][1] Popper systematically advances falsificationism: a scientific theory cannot be empirically verified, only falsified, so falsifiability becomes the boundary between science and non-science and the criterion of scientific progress. This book is the expanded English edition of

the German original *Logik der Forschung* (1934, copyright page marked 1935), published by Hutchinson of London in 1959. It is the common starting point for the later debates of Kuhn, Lakatos, and Feyerabend, and it grounds this book's posture of openly inviting refutation and writing down assertions that can be overturned.

3. I. Lakatos (1970). "Falsification and the Methodology of Scientific Research Programmes." In I. Lakatos and A. Musgrave (eds.), *Criticism and the Growth of Knowledge*. Cambridge University Press. [3] Lakatos reconciles Popper and Kuhn by proposing that the unit of evaluation is not a single theory but a "research programme": a programme has a protected hard core and an adjustable protective belt, and if it keeps predicting and delivering new facts it is progressive, otherwise degenerating. The paper grew out of a 1965 colloquium at Bedford College, London, with the collection published by Cambridge University Press in 1970. Its "progressive/degenerating" criterion provides exactly an operational methodological frame for this chapter's "theorem or pattern" dispute.

4. P. Feyerabend (1975). *Against Method: Outline of an Anarchistic Theory of Knowledge*. New Left Books. [3][1] Feyerabend, with his "epistemological anarchism," opposes any universal, fixed scientific method, maintaining that in the actual history of science "anything goes," and noting that major breakthroughs often came precisely from breaking established methodological rules. The book was first published in 1975 by New Left Books of London (later Verso). It forms the strongest opposing stance to whether cross-domain convergence is forced out by some methodological theorem: if there is no unified method at all, convergence is harder still to explain as necessary.

5. L. Laudan (1977). *Progress and Its Problems: Towards a Theory of Scientific Growth.* University of California Press. [3] Laudan argues that the measure of scientific progress is not approach to truth but "problem-solving effectiveness": how many empirical problems a theory solves and how many conceptual difficulties it creates. This frees the judgment of scientific progress from the metaphysical burden of truth and grounds it in comparable functional indicators. It bears directly on "how science progresses," and supports this chapter's restrained stance: converging on a useful pattern need not amount to converging on truth.

6. L. Laudan (1981). "A Confutation of Convergent Realism." *Philosophy of Science*, 48(1). [3][2] Laudan, with a historical list, advances the "pessimistic meta-induction": many historically successful theories that made accurate predictions had core terms later judged to refer to nothing at all (such as ether and phlogiston), so empirical success does not reliably guarantee a theory's truth. The paper appears in volume 48, pages 19 to 49. It directly challenges the theorem-like claim that "cross-domain convergence points to truth," and is one of this chapter's most crucial opposing sources.

7. J. Worrall (1989). "Structural Realism: The Best of Both Worlds?" *Dialectica*, 43(1-2). [3][2] Worrall proposes structural realism, seeking a middle path between the "no-miracles argument" and the "pessimistic meta-induction": what survives across theory change is not the description of the ontology but the mathematical structure (as Fresnel's optical equations still held after the ether was discarded). The paper appears in volume 43, issues 1 to 2, pages 99 to 124. It offers this chapter a middle reading: even if the gloss I wrap around the levers is wrong, what truly survives may be that structure itself.

8. I. Hacking (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science.* Cambridge University Press. [3][2] Hacking shifts the center of gravity in the philosophy of science from "representing" (how theory describes the world) to "intervening" (how experiment manipulates the world), proposing an experimental realism: if we can stably use an entity to intervene and to produce other phenomena ("if you can spray it, it is real"), we have reason to believe it exists. The book was published by Cambridge University Press in 1983. It offers this chapter another explanation for where the convergent pattern comes from: convergence may stem from shared experimental practice rather than from theoretical necessity.

9. W. V. O. Quine (1951). "Two Dogmas of Empiricism." *The Philosophical Review*, 60(1). [2][3] Quine attacks the two dogmas of logical empiricism: the sharp divide between the analytic and the synthetic, and reductionism. He holds that knowledge is a "web of belief" tested against experience as a whole, and that no single statement can be verified or falsified in isolation. The paper appears in volume 60, issue 1, pages 20 to 43. His confirmation holism is the philosophical bedrock of this book's central condition: no single claim can be lifted out and verified completely on its own.

10. M. Polanyi (1958). *Personal Knowledge: Towards a Post-Critical Philosophy.* University of Chicago Press. [1][4] Polanyi proposes "tacit knowledge": we know far more than we can tell, and all explicit knowledge rests on a layer of personal judgment and skill that cannot be fully formalized. Scientific cognition therefore cannot do without the scientist's personal participation and commitment. The book was published by the University of Chicago Press in 1958. It bears directly on this book's landing point: when verification cannot be exhausted, how scientists form a calibrated

belief by judgment and act upon it.

11. W. Whewell (1840). *The Philosophy of the Inductive Sciences, Founded Upon Their History.* John W. Parker. [3][2] Whewell here proposes the "consilience of inductions": when a theory induced from one class of facts turns out to explain another, originally unrelated class of facts, this unexpected convergence is a strong mark of the theory's truth. The work is in two volumes, published by John W. Parker of London in 1840. It is the earliest methodological statement of the idea that "cross-domain convergence is credible," providing the historical source for this chapter's side for "law."

12. R. K. Merton (1961). "Singletons and Multiples in Scientific Discovery: A Chapter in the Sociology of Science." *Proceedings of the American Philosophical Society*, 105(5). [1][3] Merton systematically examines the phenomenon of "multiple discovery" in the history of science: the same discovery is often made, almost simultaneously and independently, by people who knew nothing of one another, from which he argues that such multiples are not the exception but the norm of scientific discovery, with discovery depending more on the state of accumulated knowledge than on individual genius. The paper appears in volume 105, issue 5, pages 470 to 486. It is the key sociological evidence that "convergence is a strong empirical pattern," and this chapter draws on it precisely to weigh the force of independent rediscovery.

13. D. C. Dennett (1991). "Real Patterns." *The Journal of Philosophy*, 88(1). [2][3] Dennett asks: when does a pattern count as "real," rather than imposed by the observer? His criterion is compression and prediction: if describing the data as a certain pattern yields genuine information compression and supports predictions about new cases, that pattern is real. The paper appears in volume 88, issue 1, pages

27 to 51. This is the core conceptual tool of this chapter's "theorem or strong empirical pattern," and on its basis the chapter concedes that the lever decomposition has not yet predicted a move never before seen and later confirmed, a test it has still to pass.

14. N. Cartwright (1983). *How the Laws of Physics Lie.* Oxford University Press. [2][3] Cartwright argues that the fundamental laws of physics are universal precisely because they do not describe the real world truthfully: the more fundamental a law, the more idealization and approximation it needs to fit phenomena, while what truly describes concrete systems are local, phenomenological laws. The book was first published by Clarendon Press / Oxford University Press in 1983. It challenges at the root whether, behind the convergence this chapter sees, a "theorem" truly stands, or merely a tidiness at the level of models.

15. P. W. Anderson (1972). "More Is Different." *Science*, 177(4047). [2][3] Anderson opposes the reductionist "constructionist" inference: even if everything is composed of fundamental particles obeying fundamental laws, it does not follow that higher-level behavior can be derived from those laws. With each step up in scale, wholly new, self-contained regularities emerge. The paper appears in volume 177, issue 4047, pages 393 to 396 (August 4, 1972). It supports a weakened reading of this chapter: the book's convergence may be only a regularity at the scale of some "special science," not a fundamental law.

16. H. A. Simon (1962). "The Architecture of Complexity." *Proceedings of the American Philosophical Society*, 106(6). [2][3] Simon argues that complex systems that evolve stably and endure tend to be hierarchical and "nearly decomposable," with interactions within a subsystem far stronger

than those between subsystems, and uses the parable of the watchmakers to show that systems with stable intermediate parts are more easily assembled. The paper appears in volume 106, issue 6, pages 467 to 482. It gives cross-domain convergence yet another "common origin" explanation: different fields may stumble on similar structures because complex systems are subject to the same set of architectural constraints.

17. K. G. Wilson (1979). "Problems in Physics with Many Scales of Length." *Scientific American*, 241(2). [2][3] Wilson explains the renormalization group to a general readership: when a system spans many length scales (as near a critical point), it can be handled by coarse-graining step by step, thereby explaining why systems with wildly different microscopic details fall into the same "universality class" and exhibit exactly the same critical behavior. The paper appears in volume 241, issue 2, pages 158 to 179 (August 1979). It is the hardest physical evidence for this chapter's side for "law," because there the convergence of different systems to the same behavior is a provable theorem.

18. R. W. Batterman (2002). *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford University Press. [2][3] Batterman analyzes "asymptotic reasoning" in physics: many explanations (especially of universality phenomena) depend on the singular behavior that emerges when some limit is taken (such as a scale going to zero or infinity), and such explanations cannot be simply reduced to the underlying theory, residing precisely where the details vanish. The book was published by Oxford University Press (cover year usually marked 2002, some reviews dating it 2001 from the pre-publication catalog). It offers philosophical analysis for this chapter's side for "law": cross-domain convergence may hold precisely because of this

asymptotic universality, independent of microscopic detail.

19. R. Levins (1968). *Evolution in Changing Environments: Some Theoretical Explorations.* Princeton University Press. [2][3] Levins, through a series of theoretical models, explores how organisms evolve in fluctuating, uncertain environments, introducing tools such as the "fitness set" to analyze which strategy is optimal under variable conditions, throughout a modeling style that approximates complex reality with multiple simplified models. The book was published by Princeton University Press in 1968 (Monographs in Population Biology, no. 2). The multi-model, approximate modeling it represents is a biological forerunner of Wimsatt's robustness analysis, echoing this chapter's emphasis on "the convergence of many independent routes."

20. W. C. Wimsatt (1981). "Robustness, Reliability, and Overdetermination." In M. B. Brewer and B. E. Collins (eds.), *Scientific Inquiry and the Social Sciences.* Jossey-Bass. [3][2] Wimsatt systematically expounds "robustness analysis": if a conclusion can be reached again and again by many mutually independent routes of detection, derivation, or measurement, then it is more likely to be true rather than an artifact of one means, and this "overdetermination" is the key to telling real things from artifacts. The paper appears in that collection, pages 125 to 163. It is the methodological core of this chapter's "why cross-domain convergence is credible," and just the argument on which the side for "law" leans.

21. W. C. Wimsatt (2007). *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality.* Harvard University Press. [3][4] Wimsatt argues for remaking the philosophy of science into a tool fit for "limited beings": real knowers have bounded computation, are error-prone,

and are trapped in their own perspective, and so rely on heuristics, approximation, and piecewise approximation to inch toward reality rather than pursuing idealized complete rationality. The book was published by Harvard University Press in 2007. It is highly attuned to this book's theme, answering head-on how a bounded actor advances science in an unverifiable world and lives accordingly.

22. E. P. Wigner (1960). "The Unreasonable Effectiveness of Mathematics in the Natural Sciences." *Communications on Pure and Applied Mathematics*, 13(1). [2][3] Wigner marvels at a fact: mathematics developed for purely internal motives turns out, again and again, to describe natural laws with precision, a fit he calls "a wonderful gift which we neither understand nor deserve." The paper appears in volume 13, issue 1, pages 1 to 14, from the 1959 Courant Lecture. It is the prototype of the "theorem or pattern" question: is the cross-domain effectiveness of mathematics a necessity, or a vast rhyme we cannot yet explain?

23. H. Putnam (1975). *Mathematics, Matter and Method: Philosophical Papers, Volume 1.* Cambridge University Press. [2][3] This collection gathers Putnam's early papers on the philosophy of mathematics and scientific realism, among them "What is Mathematical Truth?", which gives the classic statement of the "no-miracles argument": realism is the only philosophy that does not make the success of science a miracle, for if the entities in a theory did not exist, the success of its predictions would be inexplicable. The paper appears at pages 60 to 78. It is a positive weapon for this chapter's side for "law," set squarely against Laudan's pessimistic meta-induction.

24. D. H. Wolpert and W. G. Macready (1997). "No Free Lunch Theorems for Optimization." *IEEE Transactions on Evolu-*

tionary Computation, 1(1). [2][3] Wolpert and Macready prove the "no free lunch" theorem: averaged over all possible objective functions, any two optimization algorithms have exactly the same expected performance, so there is no universally dominant algorithm, and any algorithm's advantage is bought with specific assumptions about problem structure. The paper appears in volume 1, issue 1, pages 67 to 82. It is a real theorem this chapter takes as a contrast, and one that points, of all directions, against me, reminding us that conclusions of the "unique optimal strategy" kind require very strong premises.

25. D. R. Hofstadter (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books. [2][4] Hofstadter, through Gödel's incompleteness, Escher's visual paradoxes, and Bach's canons, weaves a common motif: self-reference and "strange loops," and from it explores how mind and meaning emerge from meaningless formal hierarchies. The book was published by Basic Books in 1979 and won the Pulitzer Prize for nonfiction. It renders self-reference and recursive closure as an art, echoing this chapter's moment when a book cannot prove itself from within yet still writes itself "in the way it describes."

26. E. Nagel and J. R. Newman (1958). *Gödel's Proof*. New York University Press. [2][4] Nagel and Newman, with as little technical detail as possible, make clear to the general reader the proof strategy of Gödel's first and second incompleteness theorems: any sufficiently strong consistent formal system contains true propositions it cannot prove, and cannot prove its own consistency from within. The book was published by New York University Press in 1958. It provides a readable and accurate basis for this chapter's account of the "intrinsic limits of formal systems" and of how "a self-referential system cannot prove itself from within."

27. G. J. Chaitin (1982). "Gödel's Theorem and Information." *International Journal of Theoretical Physics*, 21(12). [2][3] Chaitin reinterprets incompleteness through algorithmic information theory: the information content implied by a formal system's axioms is finite, so it cannot prove any proposition whose complexity exceeds that content (such as that a sufficiently long random bit string is indeed random), and incompleteness is thereby reduced to an information ceiling. The paper appears in volume 21, issue 12, pages 941 to 954. It pins down the intrinsic limits of formal systems from the angle of information, providing another formal ground for the "unverifiable world."

28. M. Mitchell (2009). *Complexity: A Guided Tour*. Oxford University Press. [2][3] Mitchell surveys the science of complex systems for the general reader: from information, computation, and evolution to networks, introducing themes such as emergence and self-organization that recur across biological, computational, and social systems, and frankly admitting that the field still lacks a unified theory. The book was published by Oxford University Press in 2009. It provides a reliable introductory map to the state of research on cross-domain common patterns, and reminds us that these patterns remain, to this day, mostly empirical observation rather than theorem.

29. J. P. A. Ioannidis (2005). "Why Most Published Research Findings Are False." *PLoS Medicine*, 2(8). [3][4] Ioannidis uses a statistical model to argue that under conditions of small samples, small effects, large researcher degrees of freedom, and pervasive interests and bias, the probability that a published "positive" finding is true often falls below one half. The paper appears in volume 2, issue 8, e124 (August 2005). It is a classic alarm of metascience, reminding this chapter that a seemingly robust empirical pattern may be only the

product of systematic bias, against which the author has no reason to assume immunity.

30. Open Science Collaboration (2015). "Estimating the Reproducibility of Psychological Science." *Science*, 349(6251). [3][4] The Open Science Collaboration, coordinating over a hundred researchers, replicated one hundred published psychological studies, with only about a third of the replications yielding an effect significant and in the same direction as the original, and replication effect sizes generally smaller than originally reported. The paper appears in volume 349, issue 6251, article number aac4716. It turns the "replication crisis" from worry into quantifiable fact, providing this chapter with direct and weighty empirical data on whether a strong empirical pattern is truly robust.

Chapter 15: Knowledge Without Verification

> **Thesis:** Where epistemology comes to rest. If verification is usually unavailable, then most of what we call knowledge is knowledge without verification; and to be capable is not to "know that you are right," but to "act well, and hold a good calibration of the ways you might be wrong."

The previous chapter forced a question. If verification is usually unavailable, if even this book can offer only an unproven belief, then what exactly is that great heap of things we ordinarily call "knowledge"? This chapter sets its resting point in epistemology.

Redefining "Knowing"

In the philosophy textbooks, knowledge is "justified true belief," and best of all with a proof attached. For a bounded actor, that standard is simply out of reach on almost everything with consequences: either there is no decision procedure, or the cost explodes, or the state is hidden, or there is not enough time, or someone on the other side is working against you. By that standard, we "know" almost nothing.

So we need a different definition, one fit for a finite being. Not

"knowing that you are right," but holding a well-calibrated belief, keeping the ways you might err under control. To be capable is not to know the truth for certain, but to act well while keeping a clear-eyed sense of the ways you might be wrong. Once that turn is made, those eight moves cease to be merely an engineering toolbox; they are promoted into an epistemology, a set of methods for "how to hold a belief and act on it when the oracle never comes."

Science, the Early Prototype of This Stance

This is no new invention. Humanity's most serious institution for seeking knowledge has been doing it all along. Science never claims to verify; it says only "not yet falsified," exactly the point of Chapter 3. Science as a whole is a machine optimized for "holding belief and acting under unverifiability." The philosophers, too, said it plainly long ago. Dewey's 1929 book *The Quest for Certainty*[10] is a diagnosis in its very title: human beings spend too much energy chasing a certainty that simply does not exist in the domain of action, whereas the true function of knowledge is to guide action, not to provide insurance. James's *The Will to Believe*[9] goes further: for some things you must take a stand before the evidence is in, and in that situation choosing to believe and to wager is legitimate, not an intellectual carelessness. Polanyi's personal knowledge[8] reminds us that every "knowing" carries a personal commitment that exceeds what can be proved. Put together, these amount to a mature stance: knowledge is not a certainty waited for, but a calibrated belief put into action.

The Eight Moves, Read as an Epistemology

So we can reread those eight moves, this time not as engineering but as knowing.

The certificate is to understand one small slice thoroughly and hold the rest in doubt. Calibration is to hold graded beliefs candidly, instead of pretending to a black-and-white verdict. Redundancy is to triangulate, using several mutually independent vantage points, on something you cannot see directly. The proxy is to grasp the true target you cannot grasp by way of a tractable stand-in, while staying wary of its betrayal at the Goodhart point. Screening is to spend your limited attention where it will update you the most. The oracle is to appeal, where your own judgment falls short, to a more reliable judgment. Decay and the audit trail are the floor that keeps you "acting well": when you put a belief into practice, you make sure that if it turns out wrong, the loss can be borne, the error can be found, and it can still be corrected. Taken together, this is a usable epistemology for a finite being.

Intuition, Expertise, and the Truth About Judgment

Brought down to a particular person, how exactly does the highly skilled one manage this? This is the most concrete part of waypoint 4, and also the part most easily either mythologized or dismissed with a wave of the hand, so it has to be stated precisely.

On expert judgment, psychology has accumulated a great deal of evidence that is not always comfortable. Meehl in 1954[5] and Dawes in 1979[12] found that in many domains a simple statistical model beats the clinical intuition of experts. But another line of research offers a complementary picture. Klein's naturalistic decision making[17], Schön's "reflective practitioner"[13], the Dreyfus brothers[14], and Ericsson's research on deliberate practice[15] show that in environments with ample feedback and stable regularities, experts can develop reliable intuition, which is at bottom a well-calibrated pattern recognition honed by feedback. Gigerenzer's fast-and-frugal heuristics[16] add that simple rules work because

they have grasped the structure of the environment (ecological rationality). The most even-handed synthesis comes from Kahneman and Klein's 2009 "failure to disagree"[24]: whether intuition is worth trusting depends on the environment. In high-validity, learnable environments it is trustworthy; in low-validity, noise-filled environments it is self-deception.

This is precisely the human-shaped version of this book's epistemology. Intuition is neither magic nor garbage; it is a capacity for calibration, and calibration itself can be trained. Tetlock's Good Judgment Project[27], in a forecasting tournament run by the intelligence community, picked out a group of ordinary people called "superforecasters." They held no classified clearances and were no domain experts, yet through learnable habits, gathering evidence from multiple angles, updating in small steps, and reviewing their work severely, they reportedly pushed their forecasting accuracy to roughly thirty percent above professional analysts with access to classified intelligence. To grant this is also to grant that deliberate ignorance is sometimes rational[28], and that in the face of the genuine uncertainty of Keynes[3] and Knight[1] (what Kay and King[29] call "radical uncertainty"), positioning oneself for robustness and antifragility along Taleb's lines[26] is often wiser than pursuing precise prediction.

The Dignity of Acting Under Uncertainty

Synthesizing these observations, what emerges is a stance that is neither the paralysis of the skeptic (since nothing can be made certain, nothing counts and nothing should be done) nor the pretense of the dogmatist (enshrining a measurable number and pretending it is the unmeasurable truth). It is a third path: knowing clear-eyed what you do not know, marking that not-knowing with a scale, and acting well all the same.

There is a quiet dignity in this. To admit that verification is a

luxury is not to concede defeat; it is to take the preconditions of action seriously. A good judge does not prop himself up on certainty; he relies on not inflating his own confidence, on naming clearly the ways he might be wrong, and on the set of methods that turns that clear-eyed accounting into action.

Closing the Arc Opened in the Preface

Back to the beginning. The ancient Greeks went to Delphi to consult the oracle before setting out, and the computer scientists named that black box which instantly returns the answer an oracle too; the two share one fantasy: before you move, verify right and wrong. This book has been about the world after that fantasy breaks, and its final reply is this: the fantasy's breaking does not mean the end of knowing and acting, only that they must proceed in a different way.

For a finite being, to know was never "to wait until a proof arrived," but "to hold a calibrated belief, and to have put it into action." The oracle will not answer, but that has never, and should never, made us halt. What remains is to bring this down to a particular person, and that is the work of the afterword.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

1. F. H. Knight (1921). *Risk, Uncertainty and Profit.* Houghton Mifflin. [2][4] Knight here draws the classic distinction: between "risk," which is quantifiable and insurable, and "uncertainty," which cannot be assigned a probability, with true profit arising precisely from the latter.

When this chapter speaks of "radical uncertainty," this distinction is the source; it reminds the reader that many consequential decisions have no probability distribution to lean on at all.

2. J. M. Keynes (1921). *A Treatise on Probability.* Macmillan. [2][3] In this early work Keynes develops a logical conception of probability and introduces the notion of the "weight of evidence": our confidence in a probability judgment itself shifts with the amount of evidence. It provides a philosophical foundation for this chapter's line on "calibrated belief," showing that beyond the probability number there is a further layer of candor about the state of one's own knowledge.
3. J. M. Keynes (1937). "The General Theory of Employment." *Quarterly Journal of Economics*, 51(2), 209-223. [2][4] In this article defending the *General Theory*, Keynes admits that about many future things "we simply do not know," with no scientific basis on which to form a computable probability. It places genuine uncertainty at the center of economic behavior, and is an important forerunner of this chapter's claim that one should "act all the same in the face of an unmeasurable truth."
4. F. A. Hayek (1945). "The Use of Knowledge in Society." *American Economic Review*, 35(4), 519-530. [2][3][4] Hayek points out that the knowledge a society needs to function is never concentrated in any one place, but dispersed among countless individuals, and largely local and tacit. This article concerns how bounded actors can still coordinate their action without holding the whole picture, and it speaks directly to this chapter's situation of "no one can verify the whole, yet decisions must still be made."
5. P. E. Meehl (1954). *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence.* University of Minnesota Press. [2][4] Meehl systematically compared the predictive performance of experts' clinical judg-

ment with that of simple statistical models, concluding that the latter is often no worse than, and frequently better than, the former. This finding is the starting point for this chapter's discussion of expert intuition; it forces one to face the fact that the trustworthiness of intuition needs empirical testing rather than assumption.

6. H. A. Simon (1955). "A Behavioral Model of Rational Choice." *Quarterly Journal of Economics*, 69(1), 99-118. [2][4] Here Simon proposes "bounded rationality": real decision-makers are limited in computation, information, and time, and so "satisfice" by choosing a good-enough option rather than enumerating the optimum. This is the anthropological premise of the book's whole argument, and on it this chapter's "epistemology fit for a finite being" takes its stand.
7. H. A. Simon (1956). "Rational Choice and the Structure of the Environment." *Psychological Review*, 63(2), 129-138. [2][4] This companion piece stresses that the form of rationality depends on the structure of the environment the decision-maker is in; simple decision rules work because they fit the environment. When this chapter discusses Gigerenzer's "ecological rationality," the root of the idea can be traced back to here.
8. M. Polanyi (1958). *Personal Knowledge: Towards a Post-Critical Philosophy.* Routledge & Kegan Paul. [1][3][4] Polanyi argues that all "knowing" contains a tacit component that cannot be fully articulated, that the knower necessarily invests a personal commitment exceeding what can be proved, and that purely objective, actor-free knowledge is only an illusion. This chapter cites it to show that even the most serious pursuit of knowledge cannot escape a personal component that cannot be fully verified.
9. W. James (1897). *The Will to Believe and Other Essays in Popular Philosophy.* Longmans, Green. [3][4] James holds

that when facing choices that are momentous and forced yet underdetermined by evidence, choosing to believe and to act on that belief is legitimate, not an intellectual rashness. This chapter draws on it to show that wagering before the oracle answers can be responsible, rather than a disqualification in epistemology.

10. J. Dewey (1929). *The Quest for Certainty: A Study of the Relation of Knowledge and Action*. Minton, Balch & Company. [3][4] Dewey diagnoses humanity's fixation on certainty as a form of evasion: the true function of knowledge is to guide action and reshape situations, not to provide a once-and-for-all insurance. This book is almost the keynote of this chapter, its very title naming the fantasy the whole book sets out to dispel.
11. A. Tversky & D. Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases." *Science*, 185(4157), 1124-1131. [2][4] This foundational paper reveals that human judgment under uncertainty relies on a few heuristics (representativeness, availability, anchoring) that work most of the time but also deviate systematically from the laws of probability. When this chapter discusses where intuition is reliable and where it is not, the paper supplies the key background that "intuition makes errors with a pattern."
12. R. M. Dawes (1979). "The Robust Beauty of Improper Linear Models in Decision Making." *American Psychologist*, 34(7), 571-582. [2][4] Dawes shows that even a simple linear model with arbitrarily set weights often outpredicts expert judgment, because it uses valid cues consistently and is undisturbed by human in-the-moment fluctuation. It extends Meehl's finding and is the direct support for this chapter's section on "why simple rules are robust."
13. D. A. Schön (1983). *The Reflective Practitioner: How Professionals Think in Action*. Basic Books. [3][4] Schön proposes "reflection-in-action": skilled professionals do not rely

on applying fixed theory, but converse with the situation in the moment of practice and adjust on the fly. This chapter cites it to portray the other side of expertise, showing how reliable judgment is generated in practice with ample feedback.

14. H. L. Dreyfus & S. E. Dreyfus (1986). *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. Free Press. [4] The Dreyfus brothers propose a stage theory of skill acquisition from novice to expert, holding that the mark of higher-order expertise is moving past explicit rules into a holistic situational intuition. This chapter draws on it to show that the mature form of expertise is not greater calculation but better seeing, lending support to the idea that "intuition is a capacity that is honed."
15. K. A. Ericsson, R. Th. Krampe & C. Tesch-Römer (1993). "The Role of Deliberate Practice in the Acquisition of Expert Performance." *Psychological Review*, 100(3), 363-406. [2][4] This study argues that the key to outstanding performance is not the mere accumulation of experience but "deliberate practice," that is, effortful training with clear goals, immediate feedback, and a constant pressing toward the edge of one's ability. This chapter uses it to support a central point: reliable intuition comes from the repeated honing of feedback, not from the natural settling of time.
16. G. Gigerenzer & D. G. Goldstein (1996). "Reasoning the Fast and Frugal Way: Models of Bounded Rationality." *Psychological Review*, 103(4), 650-669. [2][4] The two authors show that fast-and-frugal heuristics such as "pick whichever one you recognize" can, in the right environment, match or even surpass complex statistical inference. When this chapter discusses "why simple rules work," this is the direct evidence, showing that less is more depends on the fit between rule and environment.
17. G. Klein (1998). *Sources of Power: How People Make Deci-*

sions. MIT Press. [2][4] Through field studies of firefighters, nurses, and other practitioners, Klein proposes "naturalistic decision making": experts often do not compare options but, through pattern recognition, quickly recognize which kind of situation the present one belongs to and what to do. This chapter cites it to present the trustworthy side of expert intuition, complementary to the statistical-model camp.

18. R. M. Hogarth (2001). *Educating Intuition*. University of Chicago Press. [2][4] Hogarth pursues the question of where intuition comes from, distinguishing "kind" from "wicked" learning environments: environments with accurate, timely feedback breed good intuition, while environments with misleading or absent feedback breed bad. This is highly consistent with this chapter's core judgment that "whether intuition is trustworthy depends on the environment."
19. G. Gigerenzer & R. Selten (Eds.) (2001). *Bounded Rationality: The Adaptive Toolbox*. MIT Press. [2][4] This collection restates bounded rationality as an "adaptive toolbox": the mind keeps a variety of simple heuristics, drawn on according to the situation, rather than pursuing a global optimum. When this chapter discusses ecological rationality, it supplies a systematized framework, gathering scattered research on heuristics into a conception of rationality.
20. G. Klein (2004). *The Power of Intuition: How to Use Your Gut Feelings to Make Better Decisions at Work*. Currency. [4] This practitioner-facing book turns Klein's research into something operable: how to accumulate experience, review decisions, and hone and scrutinize one's own intuition. For this chapter, it shows that well-calibrated intuition can not only be studied but also be deliberately cultivated.
21. P. E. Tetlock (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press. [1][2][4] Over many years tracking a large number of experts'

political and economic forecasts, Tetlock found their overall accuracy troubling, and that the more confident, grand-narrative "hedgehog" experts tended to be the less accurate. This chapter cites it both to chasten the overconfidence of experts and to lay the groundwork for the idea that "forecasting ability can be tested and trained."

22. G. Gigerenzer (2007). *Gut Feelings: The Intelligence of the Unconscious*. Viking. [4] This is Gigerenzer's popular exposition of his research: intuition is not an irrational impulse but the unconscious application of simple rules of thumb adapted to the environment, often both fast and accurate. This chapter uses it to support the view that "intuition is a form of ecological rationality."
23. N. N. Taleb (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House. [4] Taleb discusses those rare, hard-to-predict, yet enormously consequential "black swan" events, warning that people always like to concoct explanations for them after the fact while systematically underestimating their likelihood beforehand. This chapter draws on it to argue that rather than pursue precise prediction, one should position oneself for the unpredictable, echoing the later claims about robustness and antifragility.
24. D. Kahneman & G. Klein (2009). "Conditions for Intuitive Expertise: A Failure to Disagree." *American Psychologist*, 64(6), 515-526. [2][4] Belonging respectively to the "intuition is full of biases" and "expert intuition is reliable" camps, the two scholars reach consensus in this rare dialogue: whether intuition is trustworthy depends on the environment. In environments with stable regularities and ample feedback it is learnable and trustworthy; in low-validity, noise-filled environments it is self-deception. This chapter takes it as the most even-handed synthesis, the pivot of the whole section's epistemology.
25. D. Kahneman (2011). *Thinking, Fast and Slow*. Farrar,

Straus and Giroux. [2][4] Using the frame of "System 1" fast intuition and "System 2" slow reasoning, Kahneman sums up decades of research on judgment biases. This chapter draws on it to place expert intuition back into the full landscape of cognitive mechanism, reminding the reader that intuition is both a source of capacity and a source of bias.

26. N. N. Taleb (2012). *Antifragile: Things That Gain from Disorder*. Random House. [4] Taleb proposes "antifragility": beyond robustness, which merely withstands stress, some systems gain from volatility, stress, and surprise. This chapter cites it to give a positive strategy for acting under radical uncertainty, namely arranging things so that one benefits from the unpredictable rather than suffering by it.
27. P. E. Tetlock & D. Gardner (2015). *Superforecasting: The Art and Science of Prediction*. Crown. [1][4] This book reports the findings of the Good Judgment Project: a few "superforecasters" sustain accuracy higher than ordinary people, relying not on talent but on a set of learnable habits, gathering evidence from multiple angles, updating in small steps, and reviewing severely. This chapter cites it to show that calibration itself can be trained, and that forecasting is a craft that can be improved.
28. R. Hertwig & C. Engel (2016). "Homo Ignorans: Deliberately Choosing Not to Know." *Perspectives on Psychological Science*, 11(3), 359-372. [2][4] The two authors survey why and how people actively choose not to know certain information, arguing that "deliberate ignorance" is often a rational response rather than a cognitive defect. This chapter uses it to support a counterintuitive point: that sometimes not checking, not knowing, is precisely the right decision-making stance.
29. J. Kay & M. King (2020). *Radical Uncertainty: Decision-Making Beyond the Numbers*. W. W. Norton. [2][4] Continuing Knight and Keynes, Kay and King criticize the practice

of forcing all uncertainty into probability models, arguing that in the face of "radical uncertainty" one should instead ask "what is really going on here" and act through narrative and robust judgment. This book is the direct source of this chapter's phrase "radical uncertainty," and a contemporary echo of its overall keynote.

Afterword: Learning to Act Without Certainty

This book began with a structural inquiry: five faces, eight moves, four levers, a cross-domain table, and a conjecture it admits has not itself been proven. But it should end with a single person, because in the end this matter is yours.

You will spend a whole life wagering your actions on things you cannot verify. The code you write will ship with bugs you can never finish hunting; the theory you believe will demand your commitment before you have exhausted the evidence; the people you love, the people you work with, the institutions you entrust yourself to, not one of them can be verified before you place your bet, and the choices that matter most to you are precisely the least verifiable ones. The question was never whether you can be certain; you cannot. The question is whether you can act well all the same.

Everything in this book comes down to the method hidden inside that "all the same." On the small slice you can check, prove something exactly, and hold the rest in plain doubt; use several independent vantage points to triangulate what you cannot see clearly; borrow a good-enough stand-in to get a grip on what you cannot grasp, while watching closely for the moment it begins to lie to you; spend your limited attention where it can most change

your judgment; for judgments beyond your reach, ask someone more reliable; and when you bet, bet in a way that lets you survive losing, lets your errors be caught, and lets them be undone. These are not tricks. They are the whole craft of a finite person living clearly, and unparalyzed, in a world without oracles.

The ship is still in the fog. The captain never got the eyes that could see through it after all. The charts go stale, the compass drifts, the estimate of the current is forever only an estimate. Yet she changed course all the same, not because she was sure no reef lay ahead, but because staying put was every bit as much a gamble, and she had done everything she could: checked the charts, corrected the tables, left a margin, and prepared a plan for abandoning ship should she strike the reef anyway. Then she turned the rudder.

Learning to act without certainty is, in the end, learning to turn the rudder like this. The fog will not lift. That is exactly why you must learn.

References

Waypoints: 1. historical scientific judgment; 2. theoretically studied material; 3. how science progresses; 4. how to live in an unverifiable world. This section was checked source by source.

1. Aristotle (c. 4th century BCE). *Nicomachean Ethics*. [4] Aristotle here advances the concept of practical wisdom (phronesis): ethical judgment cannot be reduced to universal rules but depends on the capacity to weigh things rightly in concrete situations, and virtue is the stable character formed through repeated practice. A common English translation is that of R. C. Bartlett and S. D.

Collins (University of Chicago Press, 2011). It matters to this chapter because "acting well without certainty" is itself a form of practical wisdom: where rules give no answer, what carries you is trained judgment.

2. Epictetus (c. 125). *Enchiridion.* [4] This Stoic handbook, compiled by his pupil Arrian from the *Discourses*, turns on distinguishing what is within our control from what is not, and on drawing one's energy back to the judgments and choices that are. It echoes this chapter in this: facing a world one cannot verify and cannot control, recognizing the boundary of one's agency is the starting point for acting clearly rather than freezing.
3. Marcus Aurelius (c. 175). *Meditations.* [4] These private notes, written in Greek and titled roughly "To Himself," were never meant for publication; they record how a Stoic ruler examined himself, restrained himself, and discharged his duty amid power and impermanence. Their significance for this chapter lies in the posture they model: doing well what the present moment requires even when the outcome cannot be seen, living with uncertainty rather than seeking the shelter of certainty.
4. C. S. Peirce (1877). "The Fixation of Belief." *Popular Science Monthly*, 12, 1-15. [3][4] Peirce compares four methods by which people "fix belief": tenacity, authority, the a priori, and the method of science, and argues that only the scientific method, which appeals to an external reality and can be publicly tested and corrected, can make belief withstand the shock of doubt. It matters to this chapter because it traces the stance of "prove something exactly on the small slice, hold the rest in plain doubt" back to its source: the value of a belief lies in how it answers doubt, not in how firmly it is held.
5. W. James (1897). *The Will to Believe and Other Essays in Popular Philosophy.* Longmans, Green. [4] James ar-

gues that for questions where the evidence cannot decide, where one must nonetheless choose, and where the matter is momentous, a person has the "right to believe," because suspending judgment is itself a choice with consequences. This is precisely the central situation of this chapter: when staying put and turning the rudder are equally a gamble, declining to bet is not neutrality.

6. W. James (1907). *Pragmatism: A New Name for Some Old Ways of Thinking.* Longmans, Green. [3][4] James lays out pragmatism systematically here: the meaning and truth of an idea are gauged by what consequences it can cash out in experience and what feasible actions it can guide, not by whether it conforms to some abstract standard. It supports this chapter's view of judgment: where ultimate verification is unavailable, take "good enough, action-guiding, testable by consequences" as a practical measure of truth.
7. S. Kierkegaard (1846). *Concluding Unscientific Postscript to Philosophical Fragments.* [4] Kierkegaard, writing in Danish under the pseudonym Johannes Climacus, argues that truths bearing on existence cannot be secured by an objective system and must finally be taken up through a "leap of faith" amid uncertainty; the subjective, passionate commitment cannot be replaced by objective knowledge. Its significance for this chapter is that the most important choices are precisely the least verifiable: at some point one can only commit before the evidence is in.
8. F. H. Knight (1921). *Risk, Uncertainty and Profit.* Houghton Mifflin. [2][4] Knight draws his far-reaching distinction: "risk" is uncertainty whose probabilities are known and computable, while "uncertainty" (later often called Knightian uncertainty) is a situation in which even the probabilities cannot be estimated, and profit arises from bearing the latter. It speaks directly to this chapter's theme: the truly hard situation is not a bet with known

odds, but having to act when even the odds cannot be seen.

9. J. Dewey (1929). *The Quest for Certainty: A Study of the Relation of Knowledge and Action.* Minton, Balch. [3][4] Dewey criticizes Western philosophy for its long pursuit of an illusion of "certainty," which exalts unchanging knowledge above changeable action; he argues that knowledge is itself a process of inquiry and experiment, and its meaning lies in improving how we deal with the world. It provides this chapter's intellectual background: abandon the obsession with certainty, and treat judgment instead as a testable, revisable practice.
10. R. Niebuhr (c. 1943). *The Serenity Prayer.* [4] This widely circulated prayer asks for the serenity to accept what cannot be changed, the courage to change what can, and the wisdom to tell the two apart; its authorship and exact date are disputed, and a fairly reliable account can be found in E. Sifton (2003). *The Serenity Prayer: Faith and Politics in Times of Peace and War* (W. W. Norton). In its most distilled form it states the boundary this chapter returns to again and again: first recognize the line between what one can and cannot act upon, then spend one's effort where it can change things.
11. F. A. Hayek (1945). "The Use of Knowledge in Society." *American Economic Review*, 35(4), 519-530. [3][4] Hayek argues that the knowledge society needs is by nature dispersed, local, and hard to report centrally; no central planner can hold the whole picture, and the price mechanism is precisely the means that coordinates this dispersed knowledge and lets people decide locally for themselves. It matters to this chapter because it explains why "global verifiability" is so often a forlorn hope, and why one must triangulate an unclear whole from many local vantage points.
12. I. Berlin (1953). *The Hedgehog and the Fox: An Essay on Tolstoy's View of History.* Weidenfeld & Nicolson. [4]

Berlin, borrowing the ancient Greek fragment "the fox knows many things, but the hedgehog knows one big thing," sorts thinkers into "hedgehogs," who subsume everything under a single grand principle, and "foxes," who pursue plurality without forcing it into unity. It is useful here because it reminds us that in a complex, hard-to-verify world the fox's pluralistic judgment is often more robust than a monistic system.

13. H. A. Simon (1955). "A Behavioral Model of Rational Choice." *The Quarterly Journal of Economics*, 69(1), 99-118. [2][4] Simon here advances "bounded rationality" and "satisficing": a real decision-maker, limited in information and computation, does not search for the optimal solution but for one that is "good enough," stopping once an acceptable threshold is met. This is the theoretical bedrock of this chapter's method: attention is limited, so spend it where it matters most, seeking sufficiency rather than perfection.
14. V. E. Frankl (1959). *Man's Search for Meaning*. Beacon Press. [4] Frankl, drawing on his own experience as a concentration-camp survivor, advances "logotherapy": a person's deepest drive is the search for meaning, and even in the situations least within one's control and least verifiable in prospect, a person still keeps the freedom to choose how to face suffering. The German original was published in 1946. It matters to this chapter because it carries "standing firm even when nothing can be grasped" down to the most extreme of human experiences.
15. H.-G. Gadamer (1960). *Wahrheit und Methode: Grundzüge einer philosophischen Hermeneutik*. J. C. B. Mohr (Paul Siebeck). [4] Gadamer's founding work of philosophical hermeneutics (the English *Truth and Method* appeared in 1975) argues that understanding always proceeds from "prejudice" and within a historical situation, and that truth

cannot be reduced to a set of methodological procedures. It echoes this chapter's view of the limits of "objective verification": judgment is inseparable from standpoint, and acknowledging this is what lets one deal more candidly with the limits of one's own perspective.

16. K. R. Popper (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge & Kegan Paul. [3][4] Popper, across this collection of essays, develops his view of science: knowledge grows through bold conjecture and rigorous refutation, and a theory's value lies in being falsifiable, in daring to risk being tested. It matters to this chapter because it sets falsification up as the engine of progress: good judgment does not seek to prove itself, but seeks to expose where it might be wrong and when it begins to lie to you.
17. H. A. Simon (1969). *The Sciences of the Artificial*. MIT Press. [2][4] Simon here lays the program for "the sciences of the artificial" and for design science: every artifact (including organizations, software, and decision processes) is designed to fit goals and an environment, and design is the activity of searching for feasible solutions under constraints. It supports this chapter's view of "betting in a way that lets you survive losing, lets errors be caught and undone" as a designable practice: a good system is built to cope with uncertainty.
18. C. Argyris & D. A. Schön (1974). *Theory in Practice: Increasing Professional Effectiveness*. Jossey-Bass. [4] Argyris and Schön distinguish people's "espoused theory" from their "theory-in-use," and propose "double-loop learning": not merely correcting errors within fixed goals, but turning back to question the goals and assumptions themselves. It is useful here because it points to a habit of self-calibration: actors must be able to notice the gap between what they say and what they do, and revise accordingly.

19. A. Tversky & D. Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases." *Science*, 185(4157), 1124-1131. [2][4] This founding paper by Tversky and Kahneman reveals that human judgment under uncertainty relies on a few heuristics, such as representativeness, availability, and anchoring; these shortcuts often work, yet they systematically produce predictable biases. It matters to this chapter because it shows that our intuitive judgments about the unverifiable are not themselves wholly trustworthy, and so must be corrected through external vantage points and mechanisms.
20. D. A. Schön (1983). *The Reflective Practitioner: How Professionals Think in Action*. Basic Books. [4] Schön proposes "reflection-in-action": a skilled professional does not first think out the rules and then apply them, but thinks while doing, improvising in real-time interaction with the situation, and much professional knowledge is hard-to-articulate "tacit" knowledge. It echoes this chapter's regard for practical judgment: when one cannot see the whole and has no time to verify completely, what carries you is an on-the-spot wisdom that can correct itself in the act.
21. M. C. Nussbaum (1986). *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy*. Cambridge University Press. [4] Nussbaum argues, through Greek tragedy and philosophy, that a good life and virtue are by nature fragile, open to luck and the external world rather than self-sufficient and secure, and that trying to place the good entirely within one's control instead distorts it. It matters to this chapter because it squarely acknowledges the constitutive significance of the uncontrollable for a good life: living with fragility is itself part of ethical maturity.
22. J. C. Scott (1998). *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press. [3][4] Scott examines why so many grand

social-engineering schemes failed, and identifies the ills of "legibility" and "high modernism": top-down standardization erases the local, hard-to-encode, mētis-like practical knowledge, so that plans break loose from reality. It matters to this chapter because it warns against the superstition of a globally verifiable, quantifiable whole, and the systematic misjudgment it brings.

23. N. N. Taleb (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House. [2][4] Taleb expounds the "black swan": rare, high-impact events, explained only awkwardly after the fact, are precisely what dominate the course of history, while our models and intuitions systematically underestimate them. It matters to this chapter because it puts unpredictable, unverifiable extreme risk at the center, forcing one to rethink how to bet in such a world.
24. G. Gigerenzer (2007). *Gut Feelings: The Intelligence of the Unconscious*. Viking. [2][4] Gigerenzer argues that simple rules of thumb (heuristics) are often more accurate and more economical than complex models in a real world of incomplete information, and that so-called "intuition" is in fact an efficient shortcut adapted to its environment, a counterpoint to viewing heuristics merely as biases. It is useful here because it shows that when verification cannot be exhausted, a frugal and robust rule is often the wiser choice.
25. A. Gawande (2009). *The Checklist Manifesto: How to Get Things Right*. Metropolitan Books. [4] Gawande argues, with examples from medicine, aviation, and other fields, that in highly complex, error-prone work a simple checklist can reliably catch the crucial steps people forget or take for granted, markedly reducing failures. It directly echoes this chapter's regard for plans and mechanisms: rather than counting on making no mistakes in the moment, freeze the "just in case" beforehand into an executable procedure.
26. D. Kahneman (2011). *Thinking, Fast and Slow*. Farrar,

Straus and Giroux. [2][4] Kahneman sums up decades of his research with the distinction between the fast, intuitive "System 1" and the slow, effortful "System 2," revealing how the former produces all manner of predictable biases under uncertainty. It matters to this chapter because it systematically explains the mechanism behind the unreliability of our judgment, and so supports using deliberate, checkable methods to make up for the shortcomings of intuition.

27. N. N. Taleb (2012). *Antifragile: Things That Gain from Disorder*. Random House. [4] Taleb here advances "antifragility": some systems not only withstand shocks but gain from volatility, stress, and disorder and grow stronger; the opposite of fragile is not robust but antifragile. It matters to this chapter because it offers a positive principle for betting in an unpredictable world: seek not accurate prediction but a position where being wrong costs little and being right pays off in amplified form.
28. P. E. Tetlock & D. Gardner (2015). *Superforecasting: The Art and Science of Prediction*. Crown. [2][4] Tetlock, drawing on research from large-scale forecasting tournaments, characterizes the habits of the best-performing "superforecasters": breaking problems down, expressing themselves in probabilities rather than certainties, diligently updating in small steps as new evidence arrives, and reviewing for calibration afterward. It directly demonstrates the mode of judgment this chapter advocates: in unverifiable domains, calibrated and accountable probabilistic thinking beats feigned certainty.
29. A. Duke (2018). *Thinking in Bets: Making Smarter Decisions When You Don't Have All the Facts*. Portfolio. [4] Duke, a former professional poker player, argues for treating decisions as bets: in a world of incomplete information shot through with luck, one must judge the quality of a decision separately from the goodness of its outcome, and beware of

"resulting," praising or blaming the original choice by reading back from the result. It matters to this chapter because it offers an operational language for living with uncertainty: bet in probabilities, judge by process, rather than settling right and wrong by a single win or loss.