THE LEVERAGE ILLUSION

What Systems Thinking Promised, and What the Evidence Delivered

Abdul Jaleel Kavungal

Published by AWC (Award Winning Creations) · 2026

Contents

A Note Before We Begin 4

Part I. What We Fail to See 8

- Chapter 1. The Bathtub Test 9
- Chapter 2. Why Clever People Fail: The Anatomy of Stock-Flow Failure 23
- Chapter 3. The Beer Game: Oscillation in the Laboratory 36
- Chapter 4. Lohhausen and Tanaland: Dörner's Laboratory of Failure
- Chapter 5. Can Experience Teach Us? Learning in and About Dynamic Systems 63

Part II. Where the Promise Came From 76

- Chapter 6. The Cybernetic Moment: Feedback, Regulation and the First Promise of Control 77
- Chapter 7. General Systems and the Architecture of Complexity 89
- Chapter 8. Forrester's Wager: System Dynamics and the Idea of the Hidden Lever 99
- Chapter 9. The Soft Turn: When There Is No Single System to See 113

Part III. The Toolkit on Trial 126

- Chapter 10. Loops on Paper: What Causal Loop Diagrams Show and Hide 127
- Chapter 11. Building Models Together: What Group Model Building Delivers 138
- Chapter 12. Twelve Places to Intervene? Meadows's List and the Search for Levers 151
- Chapter 13. The Archetype Library and the Learning Organisation 169
- Chapter 14. Can Systems Thinking Be Taught, or Even Measured? 183

Part IV. Pulling the Big Levers 196

- Chapter 15. The Limits to Growth at Fifty: A Trial in Three Acts 197
- Chapter 16. Cybersyn: Viable Systems and the Dream of Real-Time Control 212
- Chapter 17. Whole Systems for Public Health: From the Foresight Map to the STOPS Trial 224
- Chapter 18. The Commons Are Not Tragic: Ostrom's Evidence and the Misperception Twist 240

Part V. Systems Thinking Without the Illusion 255
Chapter 19. After the Accident: Safety Science and the Systems Turn 256
Chapter 20. Levers Without Illusions: Capability Traps and a Practitioner's Code 270
Appendices 282
Appendix A. Evidence Grades Table 283
Appendix B. Glossary 289
Appendix C. A Reading Path 299
Appendix D. Methods Toolkit 303
Appendix E. A Note on Sources and Verification 307
References 309

A Note Before We Begin

Donella Meadows opened her best-known essay with a sentence that doubles as a confession: "Folks who do systems analysis have a great belief in 'leverage points.'" She defined those points as places in a system where "a small shift in one thing can produce big changes in everything", and she was candid about the appetite behind the idea: "We not only want to believe that there are leverage points, we want to know where they are and how to get our hands on them" (Meadows 1999). A quarter of a century later that appetite is still the main reason people pick up systems thinking. Managers want to know where to push, public health teams want to know which part of a tangled problem will move the rest, and consultants want a method that finds the push point faster than the client can. The promise that such points exist, and that a trained eye can find them, is the thread that runs through this book.

I call the promise the leverage illusion. By that I mean the belief that once you can see a system's structure, you have a reliable lever to move it. The belief has two halves, and both can fail. The first half says that we can see the structure: that with a diagram, a model or a practised habit of mind, the stocks, flows, delays and feedback loops of a situation become visible to us. The second half says that seeing the structure tells us where to act, and that acting there will produce the large, predictable effect the diagram suggests. Neither half is absurd. Each has some evidence behind it, and each has been stretched well beyond that evidence in books, courses and consulting decks. The word "illusion" is meant precisely. An illusion is a perception that feels accurate and is not, and the interesting question about any illusion is under what conditions it holds, and why.

How the book tests the belief

The book takes the belief apart one side at a time. Part I asks whether people can see systems as clearly as systems thinkers assume. It rests on the strongest evidence anywhere in the field, a long series of laboratory experiments on how people reason about accumulation, delay and feedback, and on that ground the systems thinkers turn out to have been right: most of us, experts included, misread even simple dynamic structures. Part II goes back to the primary sources to ask where the promise of levers came from, starting with the mathematics of feedback and regulation and following it as it became a management philosophy and a way of talking about societies. Part III puts the toolkit on trial: causal loop diagrams, group model building, Meadows's list of places to intervene, the archetypes of the learning-organisation movement and the attempts to teach and measure systems thinking itself. Part IV follows four real attempts to pull large levers, from the world model behind *The Limits to Growth* to the governance of shared resources, and asks what was claimed, what was done and what the evidence says happened. Part V describes what systems thinking looks like once the illusion has been set aside, using a field that has partly managed it.

The verdict that emerges is lopsided, and I think the lopsidedness is the most useful thing the book has to offer. The diagnosis that systems thinkers made about human judgment is among the best-supported findings in behavioural decision research. The cure they offered, reliable levers found by reading structure, mostly lacks evidence: the claims about training transferring to real decisions, about maps changing policy, about ranked leverage points and about systemic methods improving safety outcomes all sit at the bottom of the scale. The field's account of how people go wrong is far better evidenced than its account of how to put them right. That gap is the subject of the book, and it is also the reason the book is on the field's side. Systems thinking asks us to follow a feedback loop all the way round. The loop that runs from a method's promises to its measured results is one of those loops, and a field that takes its own advice seriously should want it followed.

Where the evidence is strong I say so plainly, and I have tried to give the researchers the credit a colleague would. The stock-flow and supply-chain experiments are solid science. Ostrom's work on the commons rests on field evidence gathered and compared across a large number of real institutions, and it holds up. Group model building reliably does some of what it claims, psychological safety has a better evidence base than the movement it grew beside, safety science has tested its own methods against each other and published the uncomfortable results, and public health has run at least one serious trial of a whole-systems intervention. Criticism in these pages is aimed at claims, and several of the sharpest criticisms were made first by the founders themselves.

How the evidence is graded

Every major claim in the book carries a grade from A to D. A grade of A means replicated experimental or quantitative findings from several independent teams, with known boundary conditions, so that the claim can be stated without heavy hedging. B means consistent evidence from several studies or a good meta-analysis, but with design limits such as correlational data, laboratory-only settings or self-report, or with rival explanations still in play. C means some supporting studies, but with small samples, single cases, before-and-after designs without controls, or conflicting results; a C claim is a plausible hypothesis. D means the claim rests mainly on theory, practitioner testimony or anecdote, has no outcome studies, or is contradicted by the best evidence available.

Grades judge a claim as it is popularly stated. Where a narrower version of the claim deserves a better grade, I give both and explain the split, so a method can earn a B for what it does in the room and a D for what it is said to do months later. Appendix A sets out the full rubric and lists every graded claim with its main supporting and contradicting sources. I make one commitment about the grades: each comes with its reasons, written into the sentence where it appears. The point of doing it that way is that you can disagree. If you think a meta-analysis deserves more weight than I gave it, or a set of case studies less, the reasons are there to argue with, and a grade you can argue with is more useful than one you are asked to trust.

Citations and figures

References are given in author-date form, such as (Sterman 1989a) or (Ostrom 1990), and each one is keyed to the References section at the end of the book. Every factual claim, figure and quotation in the book comes from a source that was checked against a live record during the research, and where a figure was confirmed only in a particular version of a study, such as a working paper, the text says so. Where a figure could not be confirmed from the source itself, I have left it out or described the study at the level the record supports, even when the number circulates widely. Examples that are hypothetical or composite are labelled as such where they appear. When a source tells a story without evidence for it, I attribute the story to the source and say that it is unsourced.

Before you turn the page

Chapter 1 begins with two short problems. They take a few minutes and need nothing beyond arithmetic. I would ask you to try them before reading any further, and to write your answers down, because the rest of the book makes more sense once you have seen, in your own handwriting, how the problems feel from the inside. Most of the people who have taken versions of these tasks were well educated, many were graduate students or health professionals, and a large share of them got the answers wrong in the same way. What those errors reveal, and how much weight they can bear, is where the argument starts.

The book opens there for a reason. A book about evidence ought to begin on its firmest ground, and the laboratory work on accumulation is the firmest ground the field has. It is also the ground on which the first half of the leverage illusion is tested most directly. If careful, educated people misread the structure of a system as small as a bathtub, then the confidence that trained eyes can read the structure of a hospital, a supply chain or a city deserves a second look, and so does every lever said to follow from that reading.

Part I. What We Fail to See

The first half of the leverage illusion is the assumption that a system's structure can be seen at all. Part I tests that assumption against the most thoroughly replicated experiments in the field, and it starts there deliberately, because a book about evidence should begin where the evidence is firmest. On this question the systems thinkers were right, and the chapters that follow say so without hedging.

Chapter 1. The Bathtub Test

Two problems to try before reading on

Take a pencil and a scrap of paper. The two problems below reproduce the logic of tasks that researchers have given to many hundreds of people in several countries. They are my own illustrative versions, with round numbers chosen for convenience, and they are not the published test items, but they ask the same kind of question. Neither needs anything beyond addition and subtraction. Please answer them before reading the rest of the chapter, and write your answers down rather than holding them in your head.

The first problem concerns a department store. The table shows how many people entered the store and how many left it during each of eight consecutive ten-minute intervals on a Saturday morning. Assume, purely for the purpose of the exercise, that 50 people were already inside when the first interval began.

Interval	People entering	People leaving
1	20	10
2	40	15
3	60	25
4	50	35
5	30	30
6	10	45
7	10	30
8	10	10

Answer four questions. In which interval did the most people enter the store? In which interval did the most people leave? At what point during the morning were the most people inside the store? At what point were the fewest people inside? Try to answer the last two by looking at the table rather than by adding up every row, and note which answer you reach first.

The second problem concerns the atmosphere. Think of the carbon dioxide in the air as water in a bathtub. Burning fossil fuels pours carbon dioxide in through the tap, and natural processes, chiefly uptake by the oceans and by plants, drain some of it out through the plughole. Suppose, for the sake of the problem, that the tap is currently running considerably faster than the drain is emptying. Now suppose that the world's emissions stop growing and are held steady, from today onward, at exactly their current rate. What happens to the amount of carbon dioxide in the atmosphere over the following decades? Does it keep rising, level off soon, or start to fall? And what would have to happen to emissions for the amount in the atmosphere to stop rising? Write down both sets of answers before you read on.

What happened when others took the test

Sweeney and Sterman (2000) published the first results of what they called a systems thinking inventory, a set of short problems that included a bathtub task and a department-store task built on the same logic as the one above. Their subjects were students at an elite business school, a population chosen for academic ability, and none of them had studied system dynamics. The tasks were deliberately elementary. They involved one stock, one inflow and one outflow, and the information needed to answer them was printed on the page.

The students understood stocks, flows and delays poorly. The design of the tasks shows where the difficulty has to lie. Questions about the flows can be answered by finding the biggest number in a column or the tallest bar in a chart, and they demand no reasoning about accumulation at all. Questions about the stock, such as when the number of people in the store was largest and when it was smallest, cannot be answered that way; they require keeping a running balance of inflow against outflow. Later work in the same

programme identified a characteristic wrong answer to questions of this kind, which is to assume that the stock follows the pattern of its flows, so that the store is fullest when most people are entering and emptiest when most are leaving (Cronin, Gonzalez and Sterman 2009). Chapter 2 examines that pattern and the evidence for it. Just as striking as the poor performance was what did not explain it. Performance did not vary systematically with the students' educational background or demographic characteristics. Whatever was going wrong seemed to be something the students shared.

Sterman and Sweeney then carried the same kind of task into climate change, where a stock-flow error has consequences beyond a classroom. In their "Cloudy skies" experiments, highly educated subjects misread the structure that links emissions to the concentration of carbon dioxide in the atmosphere, in a way that made a policy of waiting to see what happens look safer than it is (Sterman and Sweeney 2002). A later study gave MIT graduate students climate stock-flow tasks and found that most of them believed that holding emissions near their current rate would stabilise the concentration (Sterman and Sweeney 2007). As the title of that paper put it, the students' mental models of climate violated conservation of matter. If the inflow to a bathtub exceeds the outflow, the level rises, and it goes on rising for as long as that remains true, whatever the tap was doing last year. Holding emissions steady at a rate above removal keeps adding carbon dioxide to the stock every year.

If you answered the second problem by saying that the amount of carbon dioxide would level off once emissions stopped growing, you gave the answer that most of those graduate students gave. The correct answer is that it keeps rising. Stabilising the concentration requires cutting emissions until they no longer exceed the rate of removal, which, in the illustrative terms of the problem, means turning the tap down until it matches the drain. Stopping the growth of emissions only stops the rise from accelerating.

Now the department store. Most people entered during interval 3, when 60 came in, and most left during interval 6, when 45 went out. Those two answers are easy. The number inside, however, depends on the running balance. Working through the table, the store held 60 people at the end of

interval 1, 85 after interval 2, 120 after interval 3 and 135 after interval 4. During interval 5 as many people left as entered, so the number stayed at 135. It then fell to 100 after interval 6 and to 80 after interval 7, and stayed at 80 through interval 8. The store was fullest from the end of interval 4 through interval 5, two intervals after the peak in arrivals, at the moment when leaving caught up with entering. It was emptiest at the very start of the morning, with 50 people inside, which is an answer many people never consider because nothing dramatic happens in the table at that point. If you said the store was fullest in interval 3 and emptiest in interval 6, you matched each question about the stock to the most visible feature of a flow.

The one idea everything else rests on

The principle behind both answers is simple enough to state in a sentence, and the rest of this book depends on it. A stock changes only through its flows, and its level at any moment is the accumulated history of everything that has flowed in minus everything that has flowed out.

A stock is anything that accumulates: water in a tub, people in a store, carbon dioxide in the atmosphere, money in an account, patients on a waiting list, fat in a body, trust in a relationship. A flow is a rate of change, measured per unit of time: litres per minute, people per interval, tonnes per year. The net flow is the inflow minus the outflow at a given moment. When the net flow is positive, the stock rises; when it is negative, the stock falls; when it is zero, the stock stays exactly where it is, at whatever level the past has left it. The relationship is accumulation. A stock integrates its net flow over time, adding up the past so that the present level carries the whole history of the flows.

Three consequences follow, and each of them corresponds to one of the errors people made in the experiments. The first is that the stock and its flows can move in different directions. In the department store, arrivals fell sharply after interval 3 while the number of people inside kept rising, because arrivals still exceeded departures. A falling inflow is fully compatible with a rising stock. The second is that a stock peaks or bottoms out when its inflow and outflow are equal, which is usually not when either flow is at its

extreme. The third is that stocks create delay. Because a stock carries its past, it cannot jump to a new level when the flows change. It has to be filled or drained, and that takes time. The length of the delay depends on how large the stock is compared with the flows that change it, so a large stock fed by modest flows can take years to move.

Meadows used the bathtub as a teaching image for exactly this reason (Meadows 2008). Nobody needs to be taught that a bath fills when the tap runs faster than the drain, and nobody expects the water level to fall the instant the tap is turned down a little. The image works because the physical intuition is secure. The puzzle the experiments pose is why that intuition fails to transfer. The same people who would never expect a bath to empty while the tap still runs faster than the drain will say, when the problem is printed as a table of arrivals and departures or a chart of emissions, that the stock stops rising when the inflow stops rising.

It is worth asking why an intuition so secure in the bathroom should fail on paper, and one reason is that the shortcut people use is right often enough to feel reliable. In many everyday situations a stock and its inflow do move together. When a stock is small compared with its flows and drains quickly, its level settles close to whatever the inflow currently dictates. The queue at a coffee counter on a quiet morning roughly follows the rate at which customers arrive, because each customer is served and gone within minutes, and reading the queue off the arrivals gives about the right answer. The shortcut breaks down when the stock is large compared with its flows, so that it drains slowly and carries a long memory of the past, or when the inflow and outflow are set by different forces and can diverge for long periods. Those are exactly the conditions under which most organisational and policy problems operate. A heuristic that is usually harmless and occasionally disastrous is hard to unlearn, because everyday experience keeps rewarding it.

The distinction between a stock and a flow may seem too elementary to deserve this much space. It is elementary, and that is what makes the evidence important. The questions in the systems thinking inventory were not designed to catch people out with nonlinear dynamics, feedback loops or

long chains of causation. They test the most basic operation in any dynamic system. If that operation is unreliable in well-educated adults, then the more elaborate claims made for systems thinking are built on a foundation that most people do not, on the evidence, possess.

Not an artefact of one laboratory

A reasonable sceptic would ask whether all of this is a quirk of MIT. The early studies came largely from one institution and one research group, and they used students at elite schools, who might be unrepresentative in any number of ways. If the stock-flow failure were confined to one team's tasks and one population, it would be an interesting curiosity and nothing more.

The strongest answer comes from a study led by Abdel-Hamid, which took a stock-flow problem into an everyday domain and a much wider population (Abdel-Hamid et al. 2014). The task concerned weight gain over a holiday period, the kind of accumulation that almost everyone has experienced directly. It was given to 621 people in seven countries, and the sample included health professionals as well as lay members of the public. Of the lay subjects, 76 per cent failed the task, and so did 71 per cent of the health professionals. Health training, in other words, bought a difference of 5 percentage points on a problem about body weight, the very domain in which health professionals advise others for a living. The study is valuable for three reasons at once: it moves the problem into a domain everyone knows, it crosses national and cultural boundaries, and it tests the people whose expertise ought to protect them.

A second independent test came from Austria. Ossimitz gave 154 students six stock-flow tasks and found that on the task involving a budget deficit and the accumulated debt it feeds, performance was roughly at the level of chance (Ossimitz 2002). The deficit and the debt make a clean stock-flow pair: the deficit is the annual flow and the debt is the stock it adds to. A government that cuts its deficit from a large number to a smaller one is still borrowing, and its debt is still rising, only more slowly. The confusion between a shrinking deficit and a shrinking debt is exactly the confusion

between a falling inflow and a falling stock that the department-store task exposes, and it appeared in a different country, with a different research team, on a different kind of problem.

A third line of evidence concerns domain experience directly. Brunstein, Gonzalez and Kanter gave accumulation problems framed in medical terms to medical students and to other undergraduates (Brunstein, Gonzalez and Kanter 2010). The medical students did no better than the others, and their performance was unrelated to how many years of medical training they had completed. Chapter 5 returns to this study when it asks whether experience teaches; the point here is narrower. Familiarity with the subject matter of a stock-flow problem does not seem to supply the reasoning the problem requires.

The replications also have limits that belong in the record. Most of the samples in the stock-flow literature are university students in the United States and Europe, and the Abdel-Hamid study stands out precisely because it is one of the few to reach beyond them. The tasks are short, they are done on paper or on a screen, and the people taking them have little at stake in the answer. None of this makes the result doubtful, since the error appears in every population covered by the studies described here, but it does mean that the evidence speaks most directly to how people read problems of this kind, and less directly to how they behave when a real stock of their own is involved. Chapter 5 takes up that second question through experiments in which participants drawn mostly from the fisheries sector managed a simulated fish stock over which they held exclusive rights.

Put together, these studies allow a judgment that the book will not often be able to make with such confidence. I'd grade the claim that people systematically misjudge stock-flow relationships A, the highest grade in the book's rubric, because the finding has been replicated by independent teams, in several countries, with populations ranging from elite graduate students to health professionals, and on problems ranging from abstract graphs to body weight and public debt. The grade carries a boundary condition, and it should be stated in the same breath. The finding is strongest for problems presented as graphs and tables, and the size of the effect depends heavily on

how the task is presented; when the same problems are posed in words, many more people solve them. Chapter 2 examines that boundary in detail. It narrows what the result means without undoing it, because graphs and tables are the form in which most dynamic information actually reaches decision makers.

A related claim deserves a separate and lower grade. It is tempting to conclude from the Abdel-Hamid and Brunstein results that expertise and higher education offer no protection at all. I'd grade that claim B. The evidence that general and domain education help little is consistent across studies (Sweeney and Sterman 2000; Cronin et al. 2009; Brunstein et al. 2010; Abdel-Hamid et al. 2014), but there is also evidence that specific training in accumulation can improve performance, and that people with a more analytic thinking style do somewhat better. The accurate statement is that general and domain education do not protect people much, while training aimed at the stock-flow relationship itself may. Chapter 5 examines the training evidence.

The first illusion lives in our heads

The book's central idea is the leverage illusion, the belief that seeing a system's structure gives you a reliable lever to move it. The bathtub results bear on the first half of that belief, and they bear on it hard.

Systems thinking, in most of its popular forms, presents structure as something that becomes visible once you know how to look. Draw the stocks and flows, trace the loops, and the behaviour of the system will make sense; understand the behaviour, and the places to intervene will appear. That sequence assumes that the first step is the easy one. The experiments show that it is the hard one. The problems in the systems thinking inventory contain one stock and two flows, the structure is printed on the page, and the question asks only what the stock is doing. A large share of educated adults, including people whose professional training covers the domain in question, answer wrongly.

The implication is that the first leverage illusion begins before anyone draws a diagram. It begins in the ordinary habit of treating the level of something as if it moved with its rate of change. A manager who sees that new complaints have fallen this month and concludes that the complaint backlog is shrinking is reasoning the way the business-school students reasoned about the store. A minister who announces a falling deficit as if it meant falling debt is making the Austrian students' error. A public health campaign that treats a slowing rise in obesity as a turning point in the prevalence itself is reasoning about a stock as if it were a flow. None of these people lacks intelligence or information. They are applying a shortcut that serves well in many settings and fails in exactly the settings that systems thinking claims to illuminate.

This is also the point at which the book first gives the systems thinkers their due, and it should do so plainly. System dynamicists argued, before most of these experiments were run, that intuition about feedback systems is systematically unreliable (Forrester 1971a), and Sterman's catalogue of the barriers to learning in complex systems put misperceptions of accumulation and delay near the top of the list (Sterman 1994). The laboratory evidence has borne them out. On the diagnosis, the field was right. What the evidence does not show, and what the rest of the book examines, is that knowing about this weakness is enough to overcome it, or that a person trained to see stocks and flows will then see where to push.

What the error costs

The bathtub failure would matter less if the problems where it appears were rare or trivial. They are neither. Two domains show what the error costs when it leaves the laboratory, and in both the cost comes from the same source: people expect the stock to respond to a change in the flow faster and more completely than it can.

The first domain is climate. In a short piece in *Science*, Sterman argued that communicating the risks of climate change requires more than communicating the science, because people's mental models violate mass balance in a way that shapes their policy preferences (Sterman 2008). If you believe that

stabilising emissions will stabilise the concentration of carbon dioxide, then waiting to see how bad the damage becomes before acting looks prudent: the problem, on that model, will stop getting worse as soon as you stop making it worse faster. If instead you understand that the concentration keeps rising for as long as emissions exceed removal, then waiting commits you to a higher stock and therefore to larger and faster cuts later. The wait-and-see position, on this account, is less a considered judgment about costs and benefits than a consequence of a stock-flow error. Moxnes and Saysel took the problem into the laboratory, testing both the misperceptions of carbon dioxide accumulation and information policies meant to correct them (Moxnes and Saysel 2009). Their work belongs to the same programme of evidence and points to the same practical conclusion. Correcting the facts about emissions is insufficient if the listener's model of how emissions become concentration is wrong.

The second domain is the human body, and here the stock-flow structure is quantified with unusual precision. Hall and colleagues built a metabolic model of how body weight responds to changes in energy intake and validated it against data (Hall et al. 2011). Body weight is a stock; energy intake is its inflow, and energy expenditure its outflow. The model's first lesson concerns delay. When a person's energy intake changes and then stays at its new level, body weight does not jump to a new value. It moves towards a new level slowly, and it takes roughly a year to cover half the distance. Someone who changes their diet and weighs themselves a month later is looking at a stock that has barely begun to respond.

The model's second lesson concerns the size of the flows involved. Hall and colleagues estimated that a small, persistent imbalance between intake and expenditure, about 30 kilojoules a day, is enough to explain the average weight gain of the US population. That is a very small quantity of food, far too little for anyone to notice on a given day, and it is the kind of imbalance that a stock integrates, patiently, into a large result. Their third lesson runs the other way. Once the extra weight has accumulated, the model implies that keeping it requires an intake about 0.9 megajoules a day higher than before. The flow that built the stock and the flow needed to hold it at its new level are quite different quantities. Rules of thumb that treat weight as

responding immediately and in fixed proportion to a change in diet get both lessons wrong, and the Abdel-Hamid results suggest that reasoning of that kind is common among health professionals as well as the public.

The delay has a consequence for how interventions are judged. Suppose a public health programme, or a personal diet, succeeds in removing a persistent imbalance of the size Hall and colleagues describe. On their estimate, a year later the weight in question will have moved only about halfway to its new level, and the change in any given month will be small enough to vanish among ordinary fluctuations. An evaluation that looks for a clear change in the stock after a few months is therefore likely to conclude that nothing has worked, and a manager who expects the stock to follow the flow promptly will be tempted to abandon a sound programme or to pile on further measures. The first error wastes a good intervention, and the second sets up the kind of overcorrection that Chapter 3 documents in supply chains.

The two domains share a structure worth stating in general terms. When a stock is large compared with the flows that change it, three things happen together. The stock responds slowly, so early results mislead. Small persistent imbalances produce large cumulative effects, so the flows that matter may be too small to notice. And stopping the growth of an inflow does nothing to stop the growth of the stock while the inflow still exceeds the outflow, so the moment at which a problem feels as if it is coming under control is often a moment at which it is still getting worse. Each of these is a place where good intentions and bad stock-flow reasoning combine to produce a policy that arrives late, or stops early, or is judged a failure before it has had time to work.

A sketch and three questions

The practical tool that follows from this chapter is modest, and its modesty is deliberate. Chapter 2 will show that part of the stock-flow failure depends on how problems are presented, and later chapters will show that training

people to think in systems is harder than it sounds. What can be done immediately, in any meeting and on any problem, is to force the stock-flow structure onto the page before anyone draws conclusions from a trend.

The sketch takes about two minutes. Draw a box and label it with the thing that accumulates. Draw an arrow going into the box and label it with what adds to the stock, and an arrow coming out and label it with what drains it. Write the units on each: the stock in things, the flows in things per unit of time. If there are several inflows or outflows, draw them all, but resist the temptation to add loops, causes and influences at this stage. The purpose of the sketch is to make the accumulation visible, and anything that distracts from it can wait.

Two mistakes are common when people first draw the sketch, and both are worth checking for. The first is to put the measured quantity in the box because it is the number on the report, even when it is a flow: monthly sales, new cases, referrals or emissions end up labelled as if they were the thing that accumulates. The units catch this mistake quickly, since anything measured per week or per year is a flow. The second is to forget an outflow, usually the one nobody manages directly. Staff numbers have a hiring inflow that everyone watches and a resignation outflow that nobody reports until it is large; a customer base has acquisitions on the dashboard and quiet departures off it. A stock with only one flow drawn will always appear to be moving in the direction of that flow, which is the correlation shortcut reappearing in the diagram itself.

Then ask three questions, in order, and write the answers beside the sketch.

1. What is the stock? Name the quantity that accumulates and that people actually care about. It is often not the quantity being reported.
2. What are the flows? Name every inflow and outflow, and find out which of them the available data actually measure.
3. What is the net flow doing right now? Establish whether inflow currently exceeds outflow, or the reverse, and by roughly how much. That single comparison tells you whether the stock is rising or falling, whatever the direction of the latest change in either flow.

A hypothetical example shows how the check works. Imagine a regional hospital service that reports, with some satisfaction, that referrals to its orthopaedic clinic fell by 10 per cent this quarter compared with the last. The figure is a flow. The quantity patients care about is the waiting list, which is a stock. Suppose, for illustration, that the clinic receives 900 referrals a quarter after the fall and treats 800 patients a quarter. The answer to the third question is that the net flow is still positive, at 100 patients a quarter, and the waiting list is therefore still growing, only more slowly than before. The good news about referrals is genuine, but it is news about the rate of deterioration. If the service's goal is a shorter list, the relevant comparison is between the two flows, and the relevant conversation is about what would make treatments exceed referrals and by how much. If the list stands at 2,000 and the service wants to halve it within a year, it needs a surplus of treatments over referrals of about 250 a quarter, which in this illustration means treating around 1,150 patients a quarter rather than 800. The sketch does not tell the service how to get there. It tells the service what "getting there" requires, and it prevents a flow statistic from being read as a stock outcome.

The same three questions apply to a team's backlog of unresolved defects, a charity's reserves, an organisation's stock of skilled staff, the national debt or the concentration of a pollutant in a lake. In each case the check does two jobs. It separates the quantity that matters from the quantity that is measured, and it replaces the question "is the trend going the right way?" with the question that actually decides the outcome: is more flowing in than out?

The evidence behind this tool divides into two parts, and it is worth being exact about which part carries the A. The problem the tool addresses, the tendency to read a stock off its flow, is grade-A science, as this chapter has argued. The claim that a simple sketch and three questions reliably prevent the error in practice has not been tested as such, and I would not claim more for it than that it forces the relevant comparison into the open, where people who would otherwise skip it have to make it. Chapter 2 adds rules for presenting dynamic data that have some experimental support of their

own. For now the most useful single habit is also the simplest: whenever someone shows you a rate, ask what it is filling or draining, and where the level of that stock stands today.

Chapter 2. Why Clever People Fail: The Anatomy of Stock-Flow Failure

The case

Chapter 1 established the fact of the matter. Well-educated adults, including graduate students at elite universities and health professionals in seven countries, reliably misjudge simple problems about accumulation when those problems are presented as graphs or tables. A fact of that kind invites an explanation, and the explanation matters for anyone who wants to do something about it. If people fail because they cannot read graphs, the remedy is graph literacy. If they fail because the problems are abstract, the remedy is familiar examples. If they fail because they do not care, the remedy is incentives, and if they fail because the problems overload working memory, the remedy is simpler problems or better tools. Each explanation points to a different fix, and a fix aimed at the wrong cause will fail no matter how well it is delivered.

Cronin, Gonzalez and Sterman (2009) published a paper whose title stated the case as a question: why don't well-educated adults understand accumulation? The subtitle described it as a challenge to researchers, educators and citizens. The paper reported a series of experiments designed to test the obvious explanations one at a time, and its method is worth following closely, because it is the method of a good investigation. You list the suspects, you give each one a fair chance to account for the facts, and you rule out those that cannot. The suspect who survives is not thereby convicted, but the field of possibilities has narrowed, and the remaining explanations can be tested more sharply.

The method has a particular value in this literature because the obvious explanations are all comforting in the same way. Each of them locates the problem in something that can be fixed cheaply and from outside: a gap in skills, an unfamiliar setting, a lack of effort or a shortage of mental arithmetic. If any of them had turned out to be the whole story, the stock-flow failure would be a problem of education or incentives, with a known class of remedies. The experiments are interesting precisely because they closed off those comfortable routes one by one, and in doing so they pushed the explanation inward, towards the way people spontaneously frame a question about change over time.

This chapter follows that investigation, then lets the defence have its say. The defence, in this case, is a body of later research arguing that the failure depends more on the way problems are presented than the original work suggested. The verdict at the end is less tidy than a detective story usually allows. The failure is real, it is robust in the formats where it was first found, and its mechanism turns out to have more than one part.

First suspect: people cannot read graphs

The most obvious explanation is that the problems are graph-reading problems in disguise. The standard stock-flow tasks present their information as line charts or bar charts of inflows and outflows over time, and they ask questions whose answers require extracting quantities from those charts. If a large share of educated adults simply cannot read a chart accurately, the stock-flow failure would be a special case of a general weakness in graph literacy, and nothing specific to accumulation would be involved.

The suspect is plausible. Reading a chart is a skill in its own right, and there is no reason to assume that everyone who handles numbers comfortably in a table is equally comfortable when the same numbers are drawn. A task that depends on reading the height of bars at particular points could, in principle, produce errors for reasons that have nothing to do with dynamics.

Cronin and colleagues tested this suspect directly, and it did not account for the facts: the stock-flow failure was not explained by graph-reading ability (Cronin, Gonzalez and Sterman 2009). The logic of the finding is easy to see in the department-store task from Chapter 1. Anyone who answers the flow questions correctly, picking out the interval with the most arrivals and the interval with the most departures, has shown that they can read the table. The failure appeared at the next step, when the question moved from the flows to the stock. Whatever goes wrong, it goes wrong after the data have been read, when the reader has to turn a pattern of rates into a pattern of levels.

This first elimination has a practical consequence that is easy to miss. Training people to read charts more accurately would not be expected to help much with accumulation. On this evidence, graph-reading skill was not what the failing participants lacked. What they could not do, or did not do, was accumulate.

Suspects two, three and four: strangeness, indifference and overload

The second suspect is unfamiliarity. The early tasks used settings chosen for simplicity rather than relevance, such as people entering a store or water in a tub, presented as numbers and charts on a page, and it is a natural hypothesis that people reason better about content they know well. Perhaps the stock-flow failure reflects nothing more than the strangeness of reasoning about quantities that people rarely meet in daily life. On that view, set the problem in a familiar context and the failure should shrink or disappear.

The third suspect is indifference. Laboratory subjects often have little at stake, and a short paper-and-pencil task invites a quick answer. If subjects were simply not trying, the failure would be an artefact of low motivation, of interest to anyone who designs experiments but of little relevance to decision makers who care about the outcome.

Both suspects were tested, in Cronin and Gonzalez's three studies as well as in the 2009 series, and neither survived. Domain familiarity did not improve stock-flow understanding, and neither did motivation (Cronin and Gonzalez 2007). The 2009 paper reached the same conclusion about context and motivation across its own experiments (Cronin et al. 2009). The evidence from Chapter 1 fits the same pattern from another direction. The holiday weight task set the problem in one of the most familiar contexts imaginable, and 71 per cent of health professionals still failed it (Abdel-Hamid et al. 2014). Brunstein and colleagues found that medical students did no better than other undergraduates on accumulation problems framed in medical terms (Brunstein et al. 2010). Familiarity with the setting, and even professional expertise in it, did not supply the missing reasoning.

The elimination of indifference is important for a different reason. A familiar objection to laboratory research on decision making is that errors observed in cheap experiments will vanish once people care about getting things right. For the stock-flow failure, the evidence does not support the objection. Motivation, as these experiments assessed it, did not explain the error. That result gives the laboratory findings more weight as a description of how people reason when it matters, though Chapter 5 will show that experience with real stakes brings its own complications.

The fourth suspect is limited cognitive capacity. Accumulation requires a kind of running arithmetic: holding a total in mind, adjusting it by the difference between two flows at each step, and tracking the result over many steps. That is a demanding operation for working memory, and it would not be surprising if people simply lost track. If capacity were the culprit, the failure would be a failure of computation rather than of understanding, and the obvious remedy would be to hand the computation to a machine.

Cronin and colleagues found that cognitive capacity did not explain the failure either (Cronin et al. 2009). The standard tasks are already simple enough that the arithmetic is light, which is part of what makes the failure so puzzling. The department-store task in Chapter 1 does not require adding up every row to answer the question about when the store was fullest. It requires only noticing that the number inside keeps rising for as long as

arrivals exceed departures, and so finding the point at which the two become equal. That is a question about the direction of change, and it can be answered by inspection. The characteristic wrong answers in these tasks are not the scattered results of arithmetic slips. They follow a pattern, matching the stock to the most visible feature of a flow, which suggests that people are answering a different question from the one asked.

The elimination of capacity matters for anyone who hopes that better computational tools will solve the problem on their own. A spreadsheet can accumulate flows perfectly, but it cannot make someone ask for the accumulation in the first place, or notice when a chart of flows is being read as if it were a chart of levels. If the error is in the question being asked, a faster calculator does not help.

It helps to pause and consider what the four eliminations establish before turning to the suspect that remains. Taken together, they say that the stock-flow failure occurs in people who can read the display, who understand the setting, who are trying and who have the mental resources to do the calculation. That is a demanding description, and it rules out the most common excuse offered when a group of capable people misreads a trend, which is that they lacked some piece of information or skill. On this evidence the error is better described as a default: a way of answering questions about change that people fall into unless something stops them.

The eliminations also carry a caution about their own scope. Ruling out a suspect means showing that it does not explain the failure, in the sense that removing or controlling it leaves the failure in place. It does not mean that the suspect plays no part at all. Some people in these experiments presumably did misread charts, lose track of totals or answer carelessly. The claim is narrower and more useful: none of these factors is what makes the failure so common, and fixing any one of them would leave most of the failure intact.

A suspect caught talking

With the four obvious suspects set aside, attention turns to the one that Cronin and colleagues identified from the pattern of errors itself. Many well-educated people appear to use what the researchers called the correlation heuristic: they assume that a stock follows the pattern of its flows, so that the stock rises when the inflow rises and falls when the inflow falls, or peaks when the inflow peaks (Cronin et al. 2009). The heuristic produces exactly the answers seen in the department-store task. It says the store is fullest when arrivals are highest and emptiest when departures are highest. It says that carbon dioxide levels will stabilise when emissions stop rising. It says that a falling deficit means a falling debt. In each case, the level is assumed to look like the rate.

A heuristic inferred from patterns in people's answers is circumstantial evidence. The case became stronger when a later study caught the suspect in the act. Korzilius and colleagues asked participants to think aloud while solving a stock-flow task, recording what they said as they reasoned (Korzilius et al. 2014). Participants voiced the correlation heuristic directly, reasoning out loud in terms that matched the stock to the pattern of its inflow. The same protocols also surfaced other faulty strategies, which is a detail worth keeping in view. The correlation heuristic was present and audible, but it was not the only thing going wrong.

Further support comes from the training evidence, which Chapter 5 examines in detail. After a course in system dynamics, the heuristic became less common among MIT graduate students, although a minority continued to reason correlationally (Sterman 2010). And Weinhardt and colleagues found that people with an analytic thinking style did better on stock-flow tasks than those with an intuitive style, while even the analytic thinkers performed poorly (Weinhardt et al. 2015). That pattern fits the idea that the correlation heuristic is an intuitive shortcut, one that deliberate reasoning can partly override but not reliably replace.

Two rival explanations have experimental support of their own, and they prevent the case from closing. The first concerns what graphs do to attention. Cronin and Gonzalez found that graphs draw the eye to the

salient points of the flows, such as their peaks and troughs, rather than to the accumulation those flows produce (Cronin and Gonzalez 2007). On this account part of the error comes from the display itself: the most visually prominent feature of a flow chart is the high point, and people answer questions about the stock by reference to the most prominent thing in front of them. The second rival concerns scope. Fischer and Gonzalez gave 148 participants stock-flow problems and found that failure fell when the task format encouraged a global view of the whole time series, when participants had a global rather than local processing style, and even when they were primed perceptually to see wholes rather than parts (Fischer and Gonzalez 2016). That points to local attention as a cause: people who fix on individual points in a series, rather than on the series as a whole, are more likely to miss the running total.

These explanations overlap, and it is possible that they describe the same failure from different angles. A person who attends locally to the most salient point of a flow chart, and then assumes the stock looks like the flow at that point, is using the correlation heuristic, falling for graph salience and processing locally all at once. The evidence does not yet allow those contributions to be separated cleanly. One detail even cuts against a simple story. Weinhardt and colleagues found that analytic thinking style predicted better performance while global processing style did not, which sits uneasily beside Fischer and Gonzalez's finding that global processing helped (Weinhardt et al. 2015; Fischer and Gonzalez 2016). Different measures and tasks may explain the difference, but it is a reminder that the mechanism is still being mapped.

I'd grade the claim that the correlation heuristic is the main mechanism of stock-flow failure B. The heuristic is well documented as a behaviour, in the pattern of errors across experiments and now in people's own spoken reasoning, and it declines after targeted training. But at least two other mechanisms, the salience of flow peaks in graphs and the tendency to process time series locally, have experimental support, and the think-aloud work shows further faulty strategies alongside it. The correlation heuristic is the leading suspect. Calling it the sole culprit would go beyond what the evidence shows.

Cross-examination by the defence

Every investigation needs someone to argue the other side, and in this literature the defence has been conducted with care. Its central argument is that the stock-flow failure, as originally measured, is partly a product of the format in which the problems were presented. If that is right, the headline figures describe how people handle graphs of accumulation, which is a narrower thing than how they understand accumulation.

The strongest evidence for the defence comes from Fischer, Degen and Funke, who presented stock-flow problems to 107 participants in verbal form rather than as graphs (Fischer, Degen and Funke 2015). Most of them solved the problems. That is a striking result beside the failure rates reported in Chapter 1. The authors concluded that stock-flow failure is at least partly an artefact of graph formats and of the difficulty people have in constructing graphs, and I think the conclusion is fair. A person who can reason correctly about accumulation when the problem is described in words, and fails when the same problem is drawn, does not lack the concept of accumulation. What they lack is the ability, or the habit, of applying it to a particular kind of display.

Other studies push in the same direction without going as far. The global-processing results already described suggest that changing the task format can reduce failure (Fischer and Gonzalez 2016). Gonzalez and Wong ran two experiments on analogy and found that comparing problems which behave alike reduced stock-flow failure, although surface similarity between problems helped only once people knew they behaved alike (Gonzalez and Wong 2011). That qualification is important. Telling someone that the atmosphere is like a bathtub helps only if they grasp that the two share a structure; the surface resemblance alone does little. Researchers have also tested the bathtub analogy specifically as an aid to reasoning about the accumulation of carbon dioxide in the atmosphere (Guy et al. 2013), which shows how seriously the climate-communication community has taken the problem, although the details of that study's results could not be confirmed for this book.

The verbal-format result deserves a moment of scrutiny, because it can be read in two ways. On one reading it shows that people understand accumulation perfectly well, and that graphs simply get in the way. On another it shows that a well-written verbal description does part of the reasoning for the reader: a sentence that says people are still arriving faster than they leave has already performed the comparison that the graph leaves to the reader. Both readings are consistent with the evidence, and both lead to the same practical conclusion. Whatever the verbal format is doing, it helps, and it can be used deliberately. What neither reading supports is the conclusion that the stock-flow failure is unimportant, since the decisions that matter are rarely presented to people as carefully constructed sentences.

The defence has limits, and one of the most instructive comes from Newell and colleagues. They reframed the carbon dioxide accumulation problem as a matter of managing a budget. Framing it as debt improved performance; framing it as savings did not (Newell et al. 2015). If the debt frame had worked because people understand the accumulation of money better than the accumulation of gas, the savings frame should have worked too, since savings accumulate in exactly the same way. The asymmetry suggests that the improvement under the debt frame was incidental, a by-product of something about how people think about debt, rather than a genuine gain in understanding accumulation. The same study found that graphs did not reliably help. A reframing that works for accidental reasons is fragile, and it is a poor foundation for a communication strategy.

Weinhardt and colleagues supply the last piece of the cross-examination. An analytic thinking style helps, but even analytic thinkers perform poorly (Weinhardt et al. 2015). Deliberate reasoning partly rescues people from the error, and it does not rescue them fully. The review of laboratory experiments by Arango Aramburo and colleagues sets these findings in a wider context. Their review treats poor performance in dynamic tasks as the typical finding of the field's experiments, and much of the work it covers is concerned with explaining and modelling why it happens (Arango Aramburo et al. 2012).

The defence, then, has won an important concession without winning the case. The stock-flow failure depends on format. It is largest with graphs and tables, smaller with verbal descriptions, and it can be reduced by task designs that encourage a global view and by comparisons that make shared structure explicit. But the failure is robust in exactly the formats that, in my judgment, dominate professional life. Dashboards, board papers, performance reports and policy briefings commonly present dynamic information as charts and tables of rates, often without the corresponding levels, and those are the displays in which the experiments find people most reliably going wrong. This is the boundary condition Chapter 1 flagged when it graded the stock-flow claim A. The grade holds for graph and table formats. A statement such as "people cannot understand accumulation" overstates the evidence. A statement such as "people reliably fail graph-based accumulation tasks" does not.

What the verdict means for the leverage illusion

The investigation ends with a plural verdict. The failure is real and replicated. Graph-reading ability, unfamiliarity, indifference and limited capacity have all been ruled out as sufficient explanations. The correlation heuristic is the leading mechanism, with graph salience and local processing as partners whose relative importance is not settled. And a large part of the failure depends on how the problem is shown.

For the leverage illusion, this matters in a specific and slightly surprising way. Part of our blindness to structure is built into the displays we use. A chart that shows a flow without its stock invites the correlation heuristic, and a chart that highlights peaks invites people to answer stock questions by reference to those peaks. The same people, shown the same problem in words or with the accumulation made explicit, do better. That means better displays are a real lever, in the modest and testable sense that changing how dynamic information is presented changes how accurately people reason about it. It is a small lever, and the evidence for it comes from laboratory tasks rather than boardrooms. But it has a feature that most of the levers promised by systems thinking lack. It has been tested, and the tests say roughly how much it helps and where it stops helping.

There is also a warning here for systems thinking itself. The field's own favourite displays include charts of flows over time and diagrams that show influences between variables without distinguishing stocks from rates. If the evidence on format is right, those displays can reproduce the very error that systems thinking set out to correct. Chapter 10 takes up the specific case of causal loop diagrams, which hide stocks by design.

Rules for putting dynamics in front of decision makers

The practical output of this chapter is a short set of rules for presenting dynamic information to people who must make decisions from it. Each rule follows from a finding described above, and each is simple enough to apply to the next report you write or review. The support for them is laboratory evidence of the kind this book grades B to C: consistent results from several experiments, but on tasks and populations that differ from a real decision meeting, and without direct tests of the rules as a package in organisations.

1. Show the accumulation explicitly. If a decision depends on a stock, such as a waiting list, a backlog, a reserve or a concentration, put the stock on the page as its own chart or number. Do not leave the reader to infer it from the flows. This rule answers the core finding: people asked to derive a stock from its flows tend to read the stock off the flow.
2. Pair every flow chart with its stock chart. When you show arrivals and departures, show the resulting level beside them, on the same time axis. The pairing lets the reader see directly that the stock can rise while the inflow falls, which is the single most common point of confusion. It also counters the pull of graph salience, since the peak of the stock chart and the peak of the flow chart will usually sit at different points in time.
3. State the problem in words as well as in a graph. Add one or two sentences that describe the accumulation in plain language: "More patients are still being referred than treated each week, so the list is growing, although more slowly than last quarter." The verbal-format evidence suggests that this may do more for comprehension than any refinement of the chart.

4. Make the net flow visible. Where there are several flows, show the difference between total inflow and total outflow, and mark the points where it changes sign. Those are the points at which the stock turns, and they are what the reader most needs to see.
5. Encourage a view of the whole series. Present the full time series rather than a highlighted recent point, and resist designs that draw the eye to a single peak or a single month's change. The global-processing results suggest that local attention feeds the error.
6. Use analogies only with the shared structure spelled out. If you compare the problem to a bathtub or a bank account, say explicitly what plays the role of the water and the tap and the drain. The analogy evidence suggests that surface resemblance alone does little, and the framing evidence suggests that a comparison can help for the wrong reasons and fail when it is reframed.

A hypothetical illustration shows the rules working together. Imagine a local authority's housing team presenting a quarterly report to councillors. The original report contains one chart, showing the number of households applying for emergency housing each month, and the chart shows a clear decline from a peak three months earlier. The councillors, reasonably, read it as good news. A revised report following these rules would add a second chart showing the number of households currently in emergency accommodation, which, in this illustration, is still rising because placements into permanent housing are running below new applications. It would state in a sentence that the number in emergency accommodation is still growing, and it would show the monthly difference between applications and placements, making clear that the gap has narrowed but has not closed. The councillors reading the revised report are asked the same question as before, and they now have the information they need to answer it correctly.

Because the evidence for these rules comes from laboratories, you can add to it cheaply in your own setting. Take a chart that your organisation already uses to report a flow, and prepare two versions of a one-page briefing: one with the flow chart alone, as it is currently presented, and one that applies the first three rules. Give each version to a different half of a group of

colleagues and ask them the same two questions, one about the direction of the flow and one about the direction of the stock. Compare the answers. The exercise takes a few minutes, it produces a small piece of local evidence about whether your displays invite the correlation heuristic, and it tends to persuade people more effectively than any account of the published experiments, because they see the difference in their own answers.

The rules do not make the underlying reasoning easy. They make the relevant comparison visible and reduce the opportunities for a known error to operate. On the evidence reviewed in this chapter that is a worthwhile gain, and it costs almost nothing. It is also one of the few places in this book where a systems-thinking prescription rests on experiments that tested the ideas behind it, even if the rules themselves have not been tested as a package.

Chapter 3. The Beer Game: Oscillation in the Laboratory

A distribution chain on paper, 1958

In its July-August issue of 1958 the *Harvard Business Review* published an article by Forrester with a confident title: "Industrial dynamics: a major breakthrough for decision makers" (Forrester 1958). It was Forrester's first public statement of the approach that would later be called system dynamics, and at its centre was a simulation of a production and distribution system, a chain of linked stages through which orders and goods moved with delays. The idea behind it, which Chapter 8 examines as the durable core of Forrester's method, was that the behaviour of such a system, including its swings, comes largely from its own structure: from the way the stages are connected, the delays between them and the way each stage responds to what it sees.

Chapter 8 tells the story of Forrester's method and its reception. This chapter follows a narrower thread from that 1958 article: the production-distribution simulation, which anticipated the beer distribution game, a version of the same kind of chain turned into a task that people perform, and later one of the most closely examined experiments in behavioural operations. Its history is the history of a single finding tested, challenged, dismantled and reconfirmed over several decades, and it is one of the best examples in this book of a systems-thinking claim that has earned its standing.

How the game works, and a problem to try

The beer game puts players in charge of one stage each of a multi-stage supply chain, running from a retailer at one end back towards production at the other. In each round, a player receives orders from the stage downstream, ships what they can from their inventory, receives deliveries from the stage upstream, and decides how much to order. Orders and deliveries both move through the chain with delays, so that what a player orders now arrives some rounds later. The player's job is to keep their stage supplied without piling up stock it does not need. The task has almost nothing in it: one decision per round, a handful of numbers to watch, and a structure simple enough to explain in a few minutes. Players can see the chain laid out on the board in front of them.

What typically happens is what made the game worth studying. The chain produces oscillation and amplification: swings in orders and inventories that grow as they pass up the chain (Sterman 1989a). Stages run short and scramble to reorder, the scramble passes upstream in amplified form, and then the goods that were ordered in a panic arrive all at once and leave stages holding far more than they need. The swings are larger at each step back from the customer. This amplification of order variability as it moves up a chain is what later came to be called the bullwhip effect.

Before reading on, try a small problem that reproduces the logic of the decision each player faces. It is a hypothetical, with invented round numbers, and it is simpler than the published game.

You manage stock at one stage of a supply chain. Anything you order today arrives exactly three weeks from now. You aim to keep 300 units on hand. For months, demand from your customers has been steady at 100 units a week, and you have ordered 100 units a week to match. This week, demand jumps to 150 units and you have good reason to believe it will stay there. At the start of the week a delivery of 100 arrived; you then shipped the 150 your customers wanted, so your inventory has fallen from 300 to 250. Two earlier orders of 100 units each are still on their way to you. How many units do

you order this week? And if demand stays at 150 and your inventory keeps falling for the next two weeks, as it will while the old, smaller orders are still arriving, how will you respond?

Write down your order for this week and a sentence about what you would do next. The chapter returns to the problem at the end.

Fitting a rule to the players

The step that turned the beer game from a demonstration into a scientific instrument was the decision to model the players themselves. Sterman had already tested this approach in a different setting. In a 1987 study he took the decision rule that a behavioural simulation model assumed for capital investment and had human subjects make that decision inside the simulation, so that the modelled rule could be compared directly with what people actually did (Sterman 1987). The proposed heuristic explained their choices well. The method was simple in principle and powerful in practice: instead of arguing about whether a model's assumptions about decision makers were realistic, you could put decision makers in the model's place and measure.

Two years later Sterman applied the same logic to the beer game. His analysis, published in *Management Science*, showed that subjects managing the multi-stage inventory system produced oscillation and amplification that departed systematically from optimal behaviour (Sterman 1989a). The important word is "systematically". The players were not merely noisy or careless. Their decisions followed a pattern that could be captured by a simple rule of anchoring and adjustment. A player anchors on an estimate of expected demand, then adjusts the order up or down to close the gap between the inventory they want and the inventory they have. A sensible rule would also adjust for the supply line, the goods already ordered but not yet received, since those goods will close part of the inventory gap when they arrive. The players gave the supply line too little weight. When their inventory fell, they ordered to close the gap, and when the inventory was

still low next round because their earlier orders had not yet arrived, they ordered again, as if the first order had never been placed. When all those orders arrived, they found themselves with far more stock than they wanted.

It is worth being fair to the anchoring-and-adjustment rule, because it is far from foolish. Anchoring on expected demand keeps orders roughly in line with what customers want, and adjusting for the inventory gap corrects shortfalls and surpluses over time. In a system with short delays the rule works well. The trouble lies entirely in the missing term. Once there is a delay between ordering and receiving, the inventory gap stops being a reliable guide to what should be ordered, because part of that gap is already being closed by goods on the way. A rule that would be sensible with instant delivery becomes a generator of oscillation when delivery takes weeks, and the players had no way of noticing the difference from the inside, since each individual order looked reasonable given the inventory in front of them.

Sterman described the underlying problem as a misperception of feedback. The players were insensitive to the effects of their own past orders, which were working their way through the chain and would return as deliveries. They were reacting to the state of their own inventory without accounting for the feedback loop that ran from their orders, through the delay, back to that same inventory. The same year he reported a parallel result from a different task. Subjects managing capital investment in a simulated economy created costly oscillations, and the decision rules estimated from their behaviour showed the same systematic pattern of ignoring delays and the effects of their own actions (Sterman 1989b). Two tasks with different surface features produced the same underlying error.

This is the right place to name the concept that ties these results to the stock-flow failures of Chapters 1 and 2. The beer game has very few parts. Its difficulty comes from what happens between them over time: delays between action and result, and feedback from a player's own decisions back to the situation they are responding to. Researchers in the field call this dynamic complexity, to distinguish it from the complexity of a system with many components. A system can be dynamically complex with one decision maker and a few numbers, and the beer game shows how little is needed.

The supply line in the game is a stock, the goods in transit, fed by orders and drained by deliveries. Underweighting it is the correlation heuristic's close relative: attention goes to the stock that is visible and urgent, the inventory on hand, while the stock that will determine the future, the pipeline, drops out of the reasoning.

Turning up the difficulty

If misperception of feedback explains the players' behaviour, then making the feedback harder to perceive should make performance worse. Diehl and Sterman designed an experiment to test exactly that, in an inventory-management task where they could vary the length of the delay and the strength of the feedback independently (Diehl and Sterman 1995). The prediction held. Performance fell as dynamic complexity rose.

The working-paper version of the study, which is the version consulted for this book because the published article could not be accessed, gives a clearer picture of how badly performance fell. In that version, 17 MIT students enrolled, 14 undergraduates and three graduate students, and two did not finish. The design crossed three delay lengths with five levels of feedback strength, and each subject played 15 trials of 32 periods. Costs averaged 4.35 times the optimum. Performance was worse than optimal in every condition, which by itself is unremarkable, since optimal performance is a demanding benchmark. The striking result lay elsewhere. In the simple conditions subjects beat a rule of doing nothing at all, but in the most complex conditions most did worse than doing nothing, and on average the no-control rule beat the subjects in the two hardest conditions. In the working paper, about three-quarters of the shortfall came from subjects using the wrong model of the task rather than from inconsistency in applying it. Most of them ignored the supply line and undercontrolled, and the undercontrol grew with the delay and the strength of the feedback.

The finding that subjects undercontrolled may seem at odds with the beer game, where players over-order. The two fit together once the supply line is placed at the centre. A player who ignores goods in transit responds too weakly to the pipeline and too strongly to the visible gap, and the net effect

depends on the structure of the task. In the beer game, the visible gap keeps reappearing while earlier orders are in transit, so players keep ordering and end up with too much. In Diehl and Sterman's task, subjects likewise ignored the supply line, and in that structure the result was corrections too weak for the feedback they faced. In both cases the error is the same: the decision rule leaves out the part of the system that lies between action and result.

The working paper also observed that subjects did very little experimenting with their own strategies, and summarised the pattern in a single sentence: "the need to control seemed to override the ability to learn". People in charge of a system that was behaving badly kept intervening, and the intervening crowded out the kind of probing that might have taught them how the system worked. Chapter 5 returns to this theme when it asks whether experience can teach people to manage dynamic systems. For now the point is the stark one. In the hardest conditions of this experiment, active management made things worse than no management at all.

Paich and Sterman carried the investigation into a richer setting, a simulated market for a new product in which subjects played repeated trials (Paich and Sterman 1993). As the title of their paper announces, the market went through boom and bust. Performance was poor, and it worsened sharply as the feedback complexity of the market rose. The study is also one of the most important pieces of evidence on learning from experience, since subjects played repeated trials, and Chapter 5 examines what that repetition did and did not achieve.

The method running through these studies deserves credit in its own right. Sterman and his collaborators did not stop at showing that people perform badly; they wrote down the decision rule people appeared to use, estimated it from their choices and checked how much of their behaviour it explained. That turned a qualitative observation about human error into a quantitative model that others could test, extend or refute. It is one reason the finding survived the challenges that followed.

Behavioural operations takes the game apart

By the late 1990s the beer game had a well-established story attached to it: the bullwhip effect was the product of human misperception of feedback. That story then met a serious challenge from operations research. Lee, Padmanabhan and Whang published an analytical account of the bullwhip effect showing that four operational causes could amplify orders along a chain even when every agent in it behaved rationally (Lee, Padmanabhan and Whang 1997). The four causes were the way demand signals are processed into forecasts, rationing when supply is short, the batching of orders, and fluctuations in price. On this account, a manager who orders more than demand in response to a forecast update, or who inflates orders during a shortage to secure a larger share of rationed supply, may be acting perfectly sensibly given their incentives, and the amplification follows without any misperception at all.

The challenge was a fair one, and it could have undone the behavioural story. If rational responses to operational features were sufficient to produce the bullwhip, the evidence from the game might simply show players responding rationally to those features. The resolution came from a clean experimental design. Croson and Donohue ran two beer-game experiments in which they removed the operational causes Lee and his colleagues had identified, and the bullwhip effect persisted (Croson and Donohue 2006). Their decision makers consistently underweighted the supply line, just as Sterman's players had. Sharing inventory information across the chain helped somewhat, but it did not remove the underweighting.

Croson, Donohue, Katok and Sterman then pushed the design further, testing players under conditions in which demand was constant and known to everyone (Croson et al. 2014). Even there, bullwhip behaviour appeared. They traced part of it to what they called coordination risk: uncertainty about how the other players in the chain would behave. If you cannot be sure that your partners will manage their stages sensibly, it can make sense to protect yourself against their mistakes, and the protection itself becomes a source of instability. They also tested whether holding extra inventory on hand, which they called coordination stock, could serve as a remedy. The

concept matters because it shows that a behavioural bullwhip does not require anyone to misjudge customer demand. It can arise from the players' uncertainty about each other.

Other studies filled in the picture. Narayanan and Moritz ran a four-stage supply-chain experiment and found that underweighting the supply line was linked to lower scores on a test of cognitive reflection, the tendency to check an intuitive answer before giving it, and that the effect survived the standard measures intended to reduce it (Narayanan and Moritz 2015). That result connects the beer game to the finding in Chapter 2 that analytic thinkers do somewhat better on stock-flow tasks. Villa, Gonçalves and Arango examined retailers' ordering decisions under delays and found that when demand exceeded supply, retailers inflated their orders, and they quantified the drivers of that inflation (Villa, Gonçalves and Arango 2015).

The most arresting result came from Sterman and Dogan, whose paper carries the title "I'm not hoarding, I'm just stocking up before the hoarders get here." They built a version of the beer game in which hoarding was never rational: demand was constant and known, and there was no competition for supply (Sterman and Dogan 2015). Nothing in the situation justified stockpiling. Even so, 22 per cent of participants placed orders more than 25 times the level of demand. There is no operational explanation for an order of that size in a setting built to exclude one, and the authors read the result as evidence of emotional or behavioural responses to perceived scarcity. The experiment shows how far behaviour can depart from the structure of a situation once people feel exposed to shortage.

Finally, Kampmann and Sterman asked whether market institutions might correct the problem, since markets aggregate the decisions of many participants and punish poor ones. Their experiments found that markets did improve performance, but that it stayed well below optimal. Their summary is precise: "Markets moderate but do not eliminate misperceptions of feedback" (Kampmann and Sterman 2014). Chapter 18 returns to the result when it examines the management of shared resources.

Taken together, this body of work supports two claims strongly. I'd grade both A: that people misperceive feedback and delays in ways that produce oscillation in the systems they manage, and that the bullwhip effect is partly behavioural, over and above its rational operational causes. The first has been replicated across tasks, laboratories and decades. The second rests on an experimental design, removing the operational causes and still observing amplification, that is clean, simple and has been repeated. Both accounts of the bullwhip hold at once. Lee and his colleagues were right that rational responses to operational features amplify orders, and the experimentalists were right that people add amplification of their own when those features are taken away. This is one of the best-replicated findings in behavioural operations, and it deserves to be presented as solid science.

The way this debate was settled is a model of how a scientific disagreement ought to go. An analytical challenge proposed a rival explanation; experimenters designed studies that separated the two explanations; the results showed that both contributed; and later work refined the behavioural account by identifying further mechanisms, such as coordination risk, cognitive reflection and responses to scarcity. Across nearly three decades of experiments, from Sterman's analyses of 1987 and 1989 to the hoarding and market experiments published in 2014 and 2015, and more than half a century after Forrester's simulation of 1958, the core finding was attacked from several directions and came through with its essentials intact and its boundaries better understood. Few ideas associated with systems thinking have been tested this thoroughly.

The remedies on trial

A finding this robust naturally prompts the question of what fixes it, and three kinds of remedy have been tested in this literature: sharing information across the chain, shortening the delays, and training the players. The results are more mixed than the popular advice suggests.

Information sharing is a popular remedy, recommended on the reasonable-sounding grounds that if every stage could see actual customer demand and the inventories of the others, the misperceptions would have nothing to

feed on. The experimental record does not support the remedy as it is usually stated. Croson and Donohue found that sharing inventory information helped somewhat but left the supply-line underweighting in place (Croson and Donohue 2006). Steckel, Gupta and Banerji found that shorter cycle times helped, but that sharing point-of-sale data did not necessarily help and could even distract players under some patterns of demand (Steckel, Gupta and Banerji 2004). And the coordination-risk results show that bullwhip behaviour can persist even when demand is constant and known to all, which is the limiting case of perfect demand information (Croson et al. 2014).

Training has a similar record. Wu and Katok manipulated both training and communication in the beer game and found that order variability fell only when hands-on training was combined with communication among team members (Wu and Katok 2006). Training alone improved individuals' understanding of the system but not the performance of the chain. That result is worth dwelling on, because it separates two outcomes that are easily confused. Players who had been trained understood the game better, yet the chain they were part of performed no better. Understanding held by individuals did not become better collective behaviour until the players could also talk to one another.

One of these remedies differs in kind from the others. Shortening the delay does not ask players to perceive the system more accurately. It changes the system itself, so that the gap between ordering and receiving is smaller and the supply line, the thing people neglect, holds less. The Steckel result that shorter cycle times helped fits the logic of the whole programme: if the error lies in neglecting what is in transit, then reducing what is in transit reduces the damage the error can do. Changes of this sort, which alter the structure rather than trying to correct the perceptions of the people inside it, have the advantage that they keep working when people are tired, busy or under pressure. They can also be expensive, which is one reason why the cheaper remedies of sharing data and training staff are so appealing.

I'd grade the popular claim that sharing information across a supply chain cures oscillation C. Some studies find that it helps, but in the experimental record it is neither necessary nor sufficient: shorter delays help without it, it can distract, and the misperceptions can survive perfect demand information. Remedies exist, and several of them help, but they help partially, and the most effective ones work in combination.

Here the leverage illusion is tested at a small scale, and the test is revealing. The beer game is about as favourable a setting for the illusion as could be devised. The structure of the system is simple, it takes minutes to explain, and it is laid out on the board. Everyone in the game can see the whole chain. If seeing a system's structure were enough to manage it, the beer game would not oscillate. It does oscillate, reliably, with players who can see the structure. The remedies that help most either change the structure itself, as shorter delays do, or change how players coordinate, as training combined with communication does. Making the system more visible by sharing information is the least reliable of them. In the beer game, seeing the system and steering it turn out to be different skills.

Counting what is already on the way

The practical tool from this chapter comes directly from the central finding. If the characteristic error is to underweight the supply line, the remedy is to count it deliberately, every time, before placing an order or taking a corrective action. I call this supply-line accounting, and it pairs with a second habit: once a correction is on its way, wait for the delay before correcting again for the same gap.

Return to the problem from the start of the chapter. You had 250 units on hand against a target of 300, demand had risen to 150 a week, and two orders of 100 were still in transit. A natural first response, and the one that matches the anchoring-and-adjustment pattern with the supply line left out, is to order enough to cover the new demand and close the inventory gap: 150 plus 50, or 200 units. Suppose you follow that rule every week, ordering this week's demand plus whatever you are short of your target. The table below, for the same hypothetical, shows what happens.

Week	Arrives	On hand after shipping	Order placed
1	100	250	200
2	100	200	250
3	100	150	300
4	200	200	250
5	250	300	150
6	300	450	0
7	250	550	0
8	150	550	0
9	0	400	50
10	0	250	200
11	0	100	350
12	50	0	450

Demand never changes after week 1: it is 150 every week. Yet inventory falls to 150, rises to 550, falls to nothing and is heading for another surge. Your orders swing from zero to 450. Nothing in the environment caused those swings. They come entirely from ordering against the inventory gap while ignoring what is already in the pipeline, and a stage upstream receiving your orders would see a demand signal far more volatile than the steady 150 your customers actually want. In a few rows of arithmetic, the table reproduces the bullwhip.

If your own answer to the opening problem was 200 units, or something close to it, you reasoned exactly as the rule in the table does, and there is nothing careless about the answer when it is considered on its own. It covers the new demand and closes the visible gap. The difficulty only appears over the following weeks, when the inventory keeps falling because the old orders of 100 are still arriving, and the same reasonable logic calls for larger and larger orders. The table shows the cumulative result of making a sensible-looking decision repeatedly without counting the pipeline.

Now apply supply-line accounting. With a three-week delay and demand of 150, you need about 450 units in the pipeline at all times, counting the order you are about to place, simply to keep deliveries matching demand. The rule is to order enough to close two gaps: the gap between your target inventory and what you have on hand, and the gap between the pipeline you need and

what is already on order. In week 1, the inventory gap is 50 and the pipeline holds 200 against the 450 needed, so you order 50 plus 250, or 300 units. In week 2, on hand is 200 and the pipeline holds 400, so you order 100 plus 50, or 150. In week 3, on hand is 150, the pipeline holds 450, and you order 150. In week 4, the large order from week 1 arrives, on hand returns to 300, and from then on you order 150 every week and inventory stays at 300. The first order is larger than the naive one, because the pipeline needed to grow to match higher demand, but after that the system settles within the delay.

The example is deliberately idealised. It assumes you know the delay exactly, that demand is constant after a single step, and that you correct the full gap at once, which in practice would be unwise when demand is uncertain. Real ordering rules adjust partially and smooth their forecasts. What the example preserves is the principle, and the principle is general. The following steps apply to any situation in which corrective actions take time to take effect.

1. Before acting, list what is already in the pipeline: orders placed but not received, people hired but not yet started, treatments begun but not yet complete, budget committed but not yet spent.
2. Estimate how much pipeline the current rate of demand requires, which is roughly the rate multiplied by the delay.
3. Size the correction to close both gaps, the one in the stock you can see and the one in the pipeline, and no more.
4. After acting, do not correct again for the same gap until the delay has had time to pass. If the visible stock is still low next week, check whether the shortfall is new or whether it is the old one that your earlier action has not yet reached.
5. If you are one link in a chain, tell the next link what is in your pipeline. The Wu and Katok result suggests that communication among people who understand the delays is what turns individual understanding into steadier collective behaviour.

Supply-line accounting is easiest to sustain when it is built into the information people see. A report that shows inventory on hand but not goods on order invites the beer-game error, in exactly the way that Chapter 2 showed a flow chart without a stock chart invites the correlation heuristic.

Adding a pipeline column to the report, and a line showing how much pipeline current demand requires, makes the neglected stock visible at the moment of decision. It does not guarantee that people will use it, but it removes the excuse that the information was not in front of them.

The same logic applies well outside warehouses. A hypothetical hospital that recruits nurses to fill vacancies, and recruits again the next month because the wards are still short while the first cohort is still in training, will find itself overstaffed a few months later. A government that tightens policy because an indicator has not yet responded to the last tightening risks the same overshoot. In each case, the error is the one the beer game measured: acting on the visible gap as if the corrections already made did not exist.

As for how much weight this tool can bear, the problem it addresses is grade-A science, as this chapter has shown. The evidence that supply-line accounting and patience fix the problem in practice is weaker. It comes from the partial and combination-dependent successes of the remedies described above, and from the arithmetic, which shows what the rule would achieve if people followed it. Whether people do follow it under pressure is the open question, and the Diehl and Sterman working paper is a warning. Its own reading of its subjects was that the need to control overrode the ability to learn, and a rule that asks people to wait runs directly against that need. A written rule, a pipeline column on every report and a standing agreement to wait for the delay are modest defences against that need, and I suspect they are the most practical ones available.

Chapter 4. Lohhausen and Tanaland: Dörner's Laboratory of Failure

A town inside a computer

Suppose you wanted to know how people behave when they are put in charge of something genuinely complex: a system with many interacting parts, goals that conflict, consequences that arrive late and information that is incomplete. The experiments in the first three chapters of this book deliberately strip complexity away. A bathtub has one stock and two flows, and the beer game has a handful of numbers and one decision per round. Their power comes from that simplicity, since an error in a task so simple cannot be blamed on the task. But the simplicity also invites an obvious objection. Real organisations and real policy problems are not bathtubs, and it is fair to ask whether the failures seen in minimal tasks tell us anything about what happens when people face the full tangle.

A German-language research programme took the opposite approach. Instead of reducing complexity, the psychologist Dörner and his colleagues built it into the laboratory, in the form of computer simulations large enough to resemble real responsibilities. By 1980 Dörner could summarise the programme's findings for English-speaking readers in an article on the difficulties people have in dealing with complexity (Dörner 1980). One of those simulations, Lohhausen, was a town that participants were asked to govern, and the full account of it appeared in 1983 as a monograph whose subtitle announced its theme, the handling of uncertainty and complexity (Dörner et al. 1983). A later book for general readers, published in English as *The Logic of Failure*, brought Lohhausen together with other simulations from the same laboratory, including a developing region called Tanaland and a similar scenario called Moro (Dörner 1996).

This chapter reconstructs what those simulations showed and then audits them. The findings are among the most vivid in the whole literature on how people handle systems, and they are often retold outside the research literature. They deserve a fair hearing, and they also deserve the scrutiny that any vivid finding should receive. The two tasks go together. A catalogue of errors can be both illuminating and poorly measured, and working out which parts of it are which is the business of the second half of the chapter.

Mayors and development workers

A word first about the level of detail this chapter can offer. The Lohhausen monograph is in German, the English-language summaries are brief, and the popular book tells its stories without the procedural detail that a methods section would supply. The account below therefore describes the simulations at the level the sources support and no further. It does not reconstruct individual participants' decisions or the precise workings of the simulated town, because those details could not be confirmed for this book, and a dramatised version would be exactly the kind of vivid but unverified story that the audit later in the chapter warns against.

What can be said is this. In Lohhausen, participants were placed in something like the position of a town's mayor and given responsibility for governing it (Dörner et al. 1983). The simulation was complex and dynamic in the sense that runs through Dörner's work: its variables were interconnected, so that decisions in one area had consequences in others, and those consequences unfolded over time rather than appearing at once (Dörner 1980). The errors Dörner catalogued, described below, only make sense in a system of that kind. A person cannot neglect side effects in a system that has none, and cannot act ballistically in a system that reports every result immediately.

Tanaland, one of the simulations Dörner describes in *The Logic of Failure*, placed participants in the role of development workers responsible for a simulated region (Dörner 1996). It shared Lohhausen's essential features:

many interconnected variables and consequences that developed over time. The Moro scenario, which later researchers adapted for cross-cultural studies, followed the same logic (Strohschneider and Güss 1999).

The appeal of these simulations lay in their realism. They reproduced features of real responsibility that minimal tasks leave out, above all a web of interconnected variables and consequences that unfold over time. Brehmer and Dörner later defended this kind of computer-simulated microworld as a middle path between two unsatisfactory alternatives, the narrow tasks of the conventional laboratory and the uncontrolled observation of real decision makers in the field (Brehmer and Dörner 1993). A microworld, in their sense, is a simulated environment rich enough to capture the essential difficulties of a real task and controlled enough to allow systematic observation. The term has stayed in use, and the argument for it remains a good one.

It is important to be clear about what these studies were. The programme was primarily descriptive. Dörner's own popular account draws largely on his own laboratory, and it reports sample sizes and statistics sparsely (Dörner 1996). What the simulations produced was a rich body of observations about how participants behaved, organised into a set of characteristic errors. That is valuable, and it is also a different kind of evidence from the controlled comparisons of the beer-game literature.

A catalogue of errors

The errors Dörner catalogued have become part of the vocabulary of anyone who thinks about decisions in complex settings. Each is worth describing in turn, because each corresponds to a way in which a person can misread or mishandle a system's structure, and several of them echo the findings of earlier chapters.

The first is the neglect of side effects. Participants acted on the problem in front of them and did not ask what else their action would change (Dörner 1980). In a town or region built as a network, almost every intervention has consequences beyond its target, and some of those consequences feed back to undo the intended effect or to create a new problem elsewhere. A person

who treats each problem as separate will be surprised, repeatedly, by developments that were predictable from the structure. This is the same failure to account for feedback that the beer game measures in the supply line, extended to a system with many loops.

The second is what Dörner called ballistic decision making. A ballistic decision is one that, like a fired shell, is launched and then left to its fate. Participants took actions and did not check whether they had worked (Dörner 1980). They moved on to the next decision without looking back at the results of the last one, so they had no way of discovering that an intervention had failed or had produced effects they did not intend. In a system with delays this is especially damaging, because the results of an action may not appear until well after it was taken, and a decision maker who does not look for them will never connect cause to effect.

The third concerns goals. Dörner described participants shifting from goal to goal, and in some cases drifting from one theme to another without pursuing any of them to a conclusion (Dörner 1980; Dörner 1996). The later literature calls the drifting pattern thematic vagabonding. To give an illustrative picture rather than a reported case, a vagabonding mayor takes up the town's finances, then housing, then employment, then the finances again, spending a little time on each and resolving none. The opposite pattern also appears in accounts of the research: some participants retreated into a small problem they knew how to solve, and devoted their attention to it while the larger system deteriorated around them. Both patterns can be read as ways of managing the discomfort of a situation too complex to grasp. One avoids commitment by moving on; the other avoids complexity by narrowing down.

The fourth item on the list, as it is usually summarised, is goal neglect: losing sight of the goals themselves, or never making them explicit in the first place. The consequences follow from the structure of the task. Without explicit goals it is hard to judge whether any action has been a success, and in a system with many interacting parts, goals that were never compared will come into conflict at the moment when an action in pursuit of one damages another.

None of these behaviours is irrational in the narrow sense. Each economises on something scarce. Checking the results of every decision takes time and attention that could be spent on the next decision; considering every side effect requires a model of the system that the decision maker may not have; staying with one problem means neglecting others that also seem urgent; and retreating to a solvable problem at least produces a visible success. A person facing more demands than they can meet will adopt some of these economies, and in a simple, forgiving environment they will rarely be punished for it. The microworlds were built to be neither simple nor forgiving, and that is why the economies showed up as failures.

These errors have a common thread, and it links Dörner's programme to the rest of Part I. Each involves treating a dynamic, interconnected system as if it were static and separable. Neglecting side effects treats the system as if its parts were independent. Ballistic decisions treat it as if actions had immediate and certain effects. Drifting between themes and retreating into small problems are responses to a complexity that the participant has not managed to represent. The bathtub and beer-game experiments show the same underlying difficulty in its simplest form, and Dörner's simulations show what it looks like when it is multiplied across a whole system.

There is a further thread, and it connects the catalogue to the leverage illusion. On my reading, several of these errors are the behaviour of people who believe they have already understood the system. A decision maker who launches actions without checking their results must be confident about what those actions will do. One who ignores side effects must be confident that the system's parts can be handled one at a time. The errors are those of confident structure-readers, and their confidence is part of what goes wrong. If that reading is right, the first half of the leverage illusion does more than make people miss structure. It also stops them looking for the evidence that would show them what they missed.

Auditing the laboratory

The Dörner catalogue is compelling, and for a practitioner it is instantly recognisable. That is precisely why it needs auditing. A list of errors that everyone recognises from experience may be true, or it may be recognisable because it describes behaviour so general that it would appear in almost any account of people struggling with a hard problem. The question is how much weight the original studies can bear as evidence for a stable, specific pattern of failure.

The first issue is sample size and reporting. The original studies, as reported in the popular synthesis, involved small numbers of participants, and the statistics that would allow a reader to judge the strength of the findings are sparsely reported (Dörner 1996). That is not unusual for exploratory research of its period, and it does not make the observations false. It does mean that the catalogue rests more on careful observation and interpretation than on measured effects.

The second issue is reliability, and here the critique comes from inside the European tradition that Dörner helped to found. A single long microworld, such as Lohhausen, gives each participant one extended experience, and performance on that one experience may depend heavily on the particular path the simulation happened to take. Greiff, Wüstenberg and Funke argued that this makes single long microworlds unreliable as measures, and they developed an alternative, MicroDYN, which assembles many short, formally specified dynamic tasks so that performance can be measured across repeated items rather than inferred from one long run (Greiff, Wüstenberg and Funke 2012). Kluge showed another source of trouble: the difficulty of a microworld, meaning the number of relationships among its variables, strongly affects performance, knowledge and participants' confidence in their own ability, so that differences in difficulty between microworlds, or between runs of the same one, become a source of measurement error (Kluge 2008).

Funke had pursued a related programme for years. In the early 1990s he introduced microworlds built from systems of linear equations, an approach that allowed researchers to specify exactly how the variables were connected

and to vary features such as the system's own internal dynamics and the presence of side effects in a controlled way (Funke 1993). He later showed how tasks of this kind could be turned into psychometrically sound measures of how people identify and control a dynamic system (Funke 2001), and summarised the European tradition of research on complex problem solving, including the Tailorshop simulation and MicroDYN (Funke 2010). The phrase complex problem solving, usually shortened to CPS, names the field that grew out of this work: the study of how people explore, understand and control dynamic systems whose structure they do not know at the outset.

The move to short, reliable tasks did not go unchallenged, and the challenge came from Dörner himself. Writing with Funke, he argued that the psychometric reductions had lost content validity: short, tidy tasks can be measured reliably, but they risk no longer capturing what makes a problem complex in the first place (Dörner and Funke 2017). This is a genuine dilemma, and it is not one the field has resolved. The long, realistic microworld is closer to the phenomenon of interest but measures it poorly. The short, reliable task measures something well, but it may no longer be the phenomenon of interest. Osman's review of research on complex dynamic control tasks across disciplines captures the breadth of the work that has grown up in the space between, and proposes a framework centred on how people monitor and manage their own uncertainty while controlling a system (Osman 2010).

The third issue is replication. The research for this book found no pre-registered replication of the classic Dörner failure taxonomy. The catalogue has been often cited, retold and applied, but it has not been tested as a formal classification: nobody has shown that independent observers, working from a defined coding scheme, reliably identify the same errors in the same proportions across new samples.

It is worth saying what a fair test of the taxonomy would look like, because the absence of one is the main reason for caution. Independent researchers would define each failure mode in advance, in terms precise enough for trained coders to apply. They would run a new sample of participants

through one or more microworlds, record their decisions and reasoning, and have coders who did not know the hypotheses classify the behaviour. The test would ask whether the coders agreed with one another, whether the same failure modes appeared across different microworlds, and whether participants who showed a given failure mode performed worse as a result. Parts of this programme have been carried out in pieces, for example in the cross-cultural think-aloud studies described below, but not as a test of the taxonomy itself.

Putting these points together, I'd grade the claim that Dörner's microworld studies show a stable taxonomy of failure C. The failure patterns are vivid and have been observed repeatedly, and the programme's influence on the study of complex decision making is deserved. But the original studies were small, they relied on single long scenarios with known reliability problems, and the taxonomy has not been tested as a replicated classification. It is best treated as a plausible and useful description of how people often go wrong, and not as an established set of categories.

Is it just intelligence?

A second debate has run through the complex problem solving tradition for decades, and it bears directly on what systems thinking can claim to be. If people who handle microworlds well are simply people who are more intelligent, then complex problem solving is not a distinct skill, and the prospects for training it separately are correspondingly dimmer. If it is distinct, there may be something specific to teach.

The early evidence suggested distinctness. Brehmer's review of dynamic decision making reported large individual differences in how well people controlled microworlds and, at the time, no stable correlations between that performance and traditional psychological tests (Brehmer 1992). That looked like evidence that controlling a dynamic system draws on something intelligence tests do not measure.

The later evidence tells a more complicated story, and the complication is instructive. Stadler and colleagues conducted a meta-analysis of 47 studies, covering 60 samples and 13,740 participants, and found that complex

problem solving and intelligence correlate substantially, with a mean correlation of .433 (Stadler et al. 2015). The correlation was stronger for the newer, MicroDYN-type measures. The pattern suggests an explanation for the earlier null results. If single long microworlds measure performance unreliably, their correlations with anything else will be weakened, and the absence of stable correlations with intelligence in the early work may have reflected the noise in the measures more than the independence of the abilities. When measurement improved, the overlap became visible.

Overlap is not identity, and two studies with similar designs show how much the answer depends on how intelligence is measured. Wüstenberg, Greiff and Funke tested 222 university students on MicroDYN and on Raven's Advanced Progressive Matrices, a test of reasoning, and used both to predict the students' school grades (Wüstenberg, Greiff and Funke 2012). A two-dimensional model of complex problem solving fitted best, separating rule knowledge, meaning what participants learned about how the system worked, from rule application, meaning how well they controlled it. Reasoning alone explained about 12 per cent of the variance in grades, and rule knowledge alone about 10 per cent. Together they explained 18 per cent, and in the authors' words, "6% of variance is additionally explained" beyond reasoning. Rule application added nothing.

Kretzschmar, Neubert, Wüstenberg and Greiff then ran a similar study with 227 German university students, but measured intelligence far more broadly, using the Berlin Intelligence Structure test, and used two complex problem solving tools, MicroDYN and MicroFIN (Kretzschmar et al. 2016). They found that complex problem solving "represents unique variance that is not accounted for by established intelligence constructs". But its incremental validity for school grades appeared only when intelligence was measured narrowly, as figural reasoning, and not when it was measured broadly. The additional predictive power that Wüstenberg and colleagues had found against a single reasoning test disappeared against a comprehensive battery.

The numbers deserve a moment of interpretation. An additional 6 per cent of explained variance is a real increment, and in the Wüstenberg study it amounted to a third of everything the two measures explained together. The same figures are also a reminder of how much lies beyond both tests. Reasoning and rule knowledge combined left more than four-fifths of the variation in grades unexplained, and grades are themselves a limited proxy for competence in managing real systems.

The pair of studies should always be read together. The first shows that complex problem solving adds something to a narrow measure of reasoning. The second shows that it carries some variance of its own, but that the variance adds nothing to the prediction of grades once intelligence is measured broadly. I'd grade the claim that complex problem solving is distinct from intelligence B in its modest form, as overlapping but not identical, and D in its strong form, as a fully distinct ability. The meta-analysis shows substantial overlap, the Kretzschmar result shows that the evidence for distinctness depends on the measure of intelligence used, and nobody has yet shown that the unique part predicts real-world outcomes beyond what intelligence already predicts. Chapter 14 returns to this debate when it asks whether systems thinking can be measured as a skill in its own right.

Different cultures, different strategies

One more strand of the audit bears on how far the Dörner findings generalise. If the failure modes reflect something universal about human cognition, they should appear in much the same form wherever the simulations are run. If they depend partly on cultural habits of thought, the catalogue describes a particular population's errors more than a human constant.

Strohschneider and Güss gave versions of the Moro development scenario to Indian and German students and found that their problem-solving strategies differed by culture (Strohschneider and Güss 1999). Güss, Tuason and Gerhard extended the comparison to 511 students in Brazil, Germany, India, the Philippines and the United States, collecting think-aloud protocols

on two microworlds, and found cross-national differences in strategy (Güss, Tuason and Gerhard 2010). Güss and Dörner found cultural differences in strategy on a nonlinear task with time delays as well (Güss and Dörner 2011), supporting the view that cognition in complex tasks is embedded in culture.

The cross-cultural work also corrects a wider imbalance. Most of the experimental literature reviewed in Part I draws on university students in the United States and Europe, and the Güss study, with its five countries on four continents, is one of the few exceptions, alongside the seven-country weight-gain study described in Chapter 1. Claims about how people handle complex systems are, for the most part, claims about how a particular slice of people handle them, and the evidence that strategies vary by culture is a reason to hold the universal versions of those claims more loosely.

These studies do not show that the failure modes disappear in some cultures. They show that the ways people approach complex, dynamic problems vary, which limits any claim that a single catalogue of errors describes everyone. For a practitioner working across cultures, the implication is practical: the errors to watch for may differ, and a checklist built on one population's patterns should be used as a set of prompts rather than a complete list.

A pre-mortem drawn from Dörner's catalogue

The taxonomy survives the audit as a useful checklist of failure modes, even though it does not survive as a validated classification. That is enough to support a practical tool, provided it is labelled honestly. What follows is a pre-mortem for complex interventions, built on Dörner's catalogue. The evidence behind it is the C-grade evidence described above: vivid and repeated observations, without controlled tests of the taxonomy or of the checklist itself. Treat it as a heuristic for structured scepticism, not as a diagnostic instrument.

A pre-mortem asks a team to imagine that an intervention has already failed and to explain why. The questions below direct that imagination towards the specific failures Dörner observed. They work best before a decision is final, with the people who will carry it out in the room, and with someone assigned to write the answers down.

1. Side effects. For each main action, what else will it change? Which of those changes could feed back to undo the intended effect, or create a new problem somewhere we are not watching? Who outside the team would notice a side effect first?
2. Checking results. How and when will we know whether each action is working? What is the delay between the action and its first measurable effect, and have we scheduled a check after that delay rather than before it? Who is responsible for looking?
3. Goals. Have we written down what we are trying to achieve, in terms specific enough to judge success? Where do our goals conflict, and which takes priority when they do?
4. Drifting. If we look at the last few months of our attention, have we pursued any one issue to a conclusion, or have we moved from theme to theme? What would it take to stay with the most important problem until it is resolved?
5. Retreat. Are we spending our effort on the problems we know how to solve, rather than on the ones that matter most? Is there a large, uncomfortable problem that we have quietly set aside?
6. Confidence. What would we expect to see if our understanding of the system were wrong? Have we looked for it?

Consider a hypothetical example. A city council plans to reduce traffic congestion in its centre by building additional parking at the edge of the central district. Working through the checklist, the team identifies a possible side effect: easier parking could encourage more people to drive into the area, adding to congestion on the roads leading to the new car parks. They recognise that the effect on traffic would take months to appear, so they schedule a traffic count six months after opening rather than relying on the first week's impressions. They discover that two of their goals conflict, since

the local business association wants more visitors arriving by car while the council's environmental plan aims to reduce car trips, and they decide explicitly which takes precedence. Finally they ask what they would see if their model of congestion were wrong, and agree to watch traffic volumes on the approach roads as well as in the centre. None of this guarantees that the parking scheme will work. It makes the scheme's assumptions visible, sets up the checks that would reveal a failure, and puts the team in a position to learn from the result.

A pre-mortem works best when it is run as a short, disciplined exercise rather than an open discussion. Give each participant a few minutes to write down, independently, the most likely reasons the intervention failed, before anyone speaks; writing first and independently is meant to stop the group converging on the first plausible story. Then go through the six prompts in order, collecting answers and noting which failure modes the group finds hardest to rule out. The output should be a short list of checks with names and dates attached, not a record of concerns. The main risk is that the exercise turns into a ritual that is completed and filed, and the best defence is to return to the list at the first scheduled check and ask whether any of the imagined failures is beginning to appear.

The last question on the list matters most for the argument of this book. On the reading offered here, Dörner's participants failed partly because they were confident that they understood the systems they were running, and that confidence kept them from looking for the evidence that would have corrected them. A pre-mortem is a small, structured way of looking before the evidence arrives on its own.

Chapter 5. Can Experience Teach Us? Learning in and About Dynamic Systems

If people misjudge accumulation, delay and feedback as reliably as the first four chapters suggest, an obvious question follows. Surely practice helps? It would be natural to expect dynamic reasoning to improve with practice, as so many skills appear to. Experienced managers, farmers, clinicians and traders deal with stocks and delays every working day, and it would be strange if years of that experience left their judgment untouched. Perhaps the laboratory failures are the failures of novices, and the cure for them is the one that cures most novice errors: time, repetition and feedback on results.

This chapter takes that hope seriously and tests it against the evidence, one question at a time. The answer that emerges is conditional. Unstructured experience turns out to be a weak teacher of dynamics, for reasons that are themselves a matter of system structure. Structured experience, in the form of simulators, deliberate comparisons and short targeted courses, shows real gains in the laboratory and the classroom, although nobody has yet shown that those gains carry over into better decisions at work. Between the two lies a trap that matters for the argument of this book: practice can make people feel that they understand a system without changing how they understand it.

The chapter's title borrows the phrasing of Sterman's paper on learning in and about complex systems (Sterman 1994), and the two prepositions mark a useful distinction. Learning in a system means getting better at operating it: making decisions that produce better outcomes. Learning about a system means understanding its structure: knowing what accumulates, where the delays are and how the feedback runs. The two are easily confused, because

we usually infer the second from the first. A person who gets better results is assumed to understand more. Much of the evidence in this chapter concerns cases in which the inference fails.

Does practice improve results?

The cleanest test of learning from experience comes from the market experiments of Paich and Sterman, introduced in Chapter 3 for what they showed about performance under feedback complexity (Paich and Sterman 1993). Their subjects managed a new product in a simulated market over repeated trials, and the repetition is what makes the study important here. Because subjects played the same kind of task more than once, the researchers could observe what repetition, with its record of how each attempt had turned out, did to their performance and to their grasp of the market.

With experience, average results improved. Insight into the dynamics of the market did not. Whatever the subjects were learning from one trial to the next, it was something other than a better understanding of how the market worked.

The distinction between results and insight is the central idea of this chapter, and it is worth being clear about why it matters. A person can improve their results in a task through many routes that involve no understanding of its structure. They can learn to avoid the specific mistakes that hurt them last time, to be more cautious in general, or to recognise the particular pattern of this task and respond to it by rote. Each of these produces better scores on the same task. None of them transfers to a new task with a different surface and the same underlying structure, and none of them would help a manager facing a market that behaves differently from the one they practised on. Better results on a practised task are therefore weak evidence that someone has learned about the system, and they can be actively misleading if the person takes them as proof of understanding.

What happens when the resource is your own?

A different line of research tested experience in a setting where the participants' familiarity with the domain could hardly have been greater. Moxnes recruited 83 participants, most of them from the Norwegian fisheries sector, and gave each one a simulated fish stock to manage (Moxnes 1998). The design had a crucial feature. Each participant held exclusive property rights over their stock. There were no other fishers competing for the same fish, no incentive to take a share before someone else did, and no tragedy of the commons in the usual sense. Any overuse would have to come from the participant's own misjudgement of the resource.

Overuse came anyway. Participants invested in more fishing capacity than the stock could support, with average overcapacity of 60 per cent, and they drew the stock down to about 15 per cent below its optimum. These were mostly people from the fisheries sector, managing a resource they owned outright, and they overinvested in exactly the way that the commons literature usually attributes to competition among users.

Moxnes then ran a series of four laboratory experiments on renewable resources, each designed to rule out the commons problem, and summarised them together (Moxnes 2000). Subjects still overinvested and overharvested. Moxnes attributed the pattern to misperceptions of the resource's stocks and flows and of the nonlinear relationships that govern its growth. The fish stock is a stock in exactly the sense of Chapter 1, and investment in fishing capacity is a supply line of the kind that Chapter 3 showed people neglecting. In a resource system of this kind, capacity built today keeps harvesting for years, and a manager who judges how much capacity to add by looking at today's catch, while the stock is quietly being drawn down, will build too much.

The reindeer experiments took the simplification further. Moxnes built simulators of reindeer grazing on rangeland, reduced to one stock and two flows, and gave participants repeated opportunities to manage them with feedback on their outcomes (Moxnes 2004). The misperceptions persisted, even with repeated outcome feedback. When the simulator was made slightly more complex, with two stocks instead of one, performance got

worse. A structure as simple as the bathtub in Chapter 1, managed repeatedly with information about how things were going, did not teach participants to manage it well.

These experiments support the claim that resource overuse can arise from misperception of dynamics, even where there is no commons dilemma at all. I'd grade that claim B. The programme is striking and internally consistent, the designs rule out the obvious alternative explanation, and the research for this book found no study contradicting it. But it comes largely from one research group, the samples are modest, and replication outside Moxnes's programme is limited. Chapter 18 returns to these results as a twist in the story of the commons, where they complicate the view that overuse is always a problem of institutions.

Does professional training protect?

Most of the fisheries participants brought experience of the domain, and it did not save them. The evidence on professional training points the same way. Brunstein and colleagues found that medical students did no better than other undergraduates on accumulation problems framed in medical terms, and that performance was unrelated to the number of years of medical training they had completed (Brunstein et al. 2010). Chapter 1 described the seven-country study in which health professionals failed a weight-gain task at nearly the same rate as lay people.

The medical result is instructive because medical training is long, and much of what it teaches concerns quantities in the body that rise and fall over time. If domain experience of that kind were enough to build stock-flow reasoning, it should have shown up in these results. It did not. Knowing a great deal about a domain, and practising within it, is evidently different from acquiring the general skill of reasoning about accumulation that the domain's problems require.

Why is experience such a poor teacher?

The pessimistic results have a theoretical explanation, and it comes from inside system dynamics. Sterman set it out in a long paper on learning in and about complex systems, which treats learning itself as a feedback process (Sterman 1994). A person acts, observes the results, revises their understanding and acts again. Learning happens when that loop works: when the results of actions are visible, arrive soon enough to be connected to the actions that caused them, and are clear enough to distinguish the effect of one's own decisions from everything else going on.

In complex systems, Sterman argued, the learning loop is broken in several places at once. Dynamic complexity means that the effects of an action are spread out in time and across the system, so they are hard to trace back. Information is poor, often arriving late, aggregated or distorted. Outcomes are confounded, because many things change at once and the effect of any one decision is hidden among them. People's mental models already contain the misperceptions documented in the experiments, so they interpret feedback through a distorted lens. And defensive routines in organisations make it hard to admit and examine failures. Each of these weakens the feedback that learning needs, and together they explain why experience in a complex system can accumulate for years without improving understanding. Sterman's proposed remedy was the use of virtual worlds, simulated environments in which the learning loop could be repaired by making consequences visible and quick.

Sterman later carried the argument into public health, arguing that policies meet resistance and fail partly because learning feedback in complex systems is weak, slow and confounded, and that simulation should be used alongside the conventional methods of evidence-based practice (Sterman 2006). The argument applies with equal force to management. An organisation that changes a policy and sees performance improve a year later has no easy way of knowing whether the policy caused the improvement, whether something else did, or whether the improvement would have been larger without the policy.

Rahmandad, Repenning and Sterman examined one link in the chain in detail, using simulation to study how learning heuristics perform in a resource-allocation task when the payoff to an action is delayed (Rahmandad, Repenning and Sterman 2009). When the decision maker misjudged the length of the delay, learning slowed sharply or never found the best allocation at all. The study is a simulation of learning rules rather than an experiment with people, so it shows what follows from its assumptions rather than what people actually do. But its message fits the experimental evidence closely. A delay that the learner does not account for can stop learning entirely, because the learner attributes results to the wrong actions and draws the wrong conclusions from them. The Diehl and Sterman subjects in Chapter 3, whose need to control overrode their ability to learn, were caught in a version of the same trap.

A hypothetical illustration shows how these breaks combine in an ordinary organisation. Suppose a software company decides to cut its backlog of customer support tickets by hiring more support staff. New staff take several months to recruit and several more to become fully productive, so for the first half-year the backlog keeps growing, and experienced staff spend time training newcomers instead of closing tickets. Meanwhile, a product release happens to reduce the number of new tickets. When the backlog finally starts to fall, managers credit the hiring, though the release may have done most of the work, and they have no easy way of telling. The lesson they carry into the next decision, that hiring cuts backlogs within a year, may be wrong in either direction. Every element of Sterman's list is present: a delay, poor and aggregated information, a confounding change and a mental model that shapes how the result is read.

On the claim that experience and outcome feedback alone do not fix misperceptions of dynamics, I'd give a grade of B. The Paich and Sterman market experiments, the Moxnes reindeer studies and the medical-training results point the same way, and there is a coherent theory explaining why. The grade is not higher because the evidence is laboratory-based and because the contrary evidence, reviewed next, shows that some kinds of experience do help. The accurate claim is conditional: unstructured experience is a weak teacher of dynamics.

What kinds of experience do help?

The optimistic evidence concerns experience that has been designed to teach. The best of it comes from studies that changed how people encountered a dynamic problem and measured whether their understanding improved.

Kopainsky and Sawicka added an interactive simulator to the written description of a natural-resource management task, and found in laboratory experiments that the simulator improved both performance and understanding of the system (Kopainsky and Sawicka 2011). The combination matters. Performance alone could improve for the reasons discussed earlier, without any gain in insight. Here understanding improved as well, which suggests that the simulator gave participants something that the text alone did not, most plausibly a way of seeing the consequences of their decisions play out over time.

Dutt and Gonzalez studied whether repeated decisions with feedback in a climate simulator reduce misconceptions about climate change, and the title of their paper reports that they do (Dutt and Gonzalez 2012). The details of their results could not be confirmed for this book, so I cite the study only for its topic, but it belongs with the evidence that experience, when it is structured to show the accumulation, can help. Chapter 2 described the related work of Gonzalez and Wong, which found that comparing problems that behave alike reduces stock-flow failure (Gonzalez and Wong 2011); a deliberate comparison is a form of structured experience, and its success depends on making the shared structure explicit. Gonzalez, Vanyukov and Martin set out the methodological case for microworlds as tools for studying dynamic decision making, along with their defining features and the trade-offs they make against realism (Gonzalez, Vanyukov and Martin 2005), and the same trade-offs apply when microworlds are used for teaching.

There is also a long practitioner literature on using models and simulations to help organisations learn, collected for example in the edited volume by Morecroft and Sterman (1994). Its cases are rich and often persuasive, but they are accounts written by the people who designed and ran the interventions, and they belong to the category of testimony rather than measured

outcome. They show what practitioners believe their simulators achieve, which is useful to know, and they leave open how much those beliefs would survive a controlled comparison.

The classroom evidence comes from short courses. The most substantial study is Sterman's own evaluation of an introductory system dynamics course taken by MIT graduate students (Sterman 2010). He tested the students before and after the course on accumulation problems and found a large and statistically significant improvement in their reasoning, together with a decline in use of the correlation heuristic. A minority, however, still reasoned correlationally after the course. The study has clear limits, which Sterman's sample makes obvious: MIT graduate students are a highly selected group, and there was no randomised control group, so the improvement cannot be cleanly separated from other things that happened to the students during the course.

Other courses report similar patterns with similar limitations. Kainz and Ossimitz tested students before and after a 90-minute crash course on stock-flow concepts and reported improvement, with no control group noted (Kainz and Ossimitz 2002). Pala and Vennix designed a pre-test and post-test around a 13-week introductory system dynamics course, using three tasks from the systems thinking inventory: the department-store task, a manufacturing task and a carbon dioxide task concerned with zero emissions (Pala and Vennix 2005). Their results could not be confirmed from the primary paper for this book, so I describe only the design, which is itself informative: it measured learning on the same kinds of tasks that first revealed the failure.

Simulation games designed for learning form a final strand. WORLD CLIMATE is a negotiation role-play for climate learning built on C-ROADS, a small climate policy simulator that links emissions pledges to concentrations and warming (Sterman et al. 2015). Chapter 20 discusses C-ROADS as an example of a small, transparent model. The WORLD CLIMATE paper reports evaluation evidence, but its effect sizes could not be confirmed for this book, and I state none. The exercise matters here as an

example of the kind of structured experience the evidence favours: it puts participants in the position of making decisions and seeing their accumulated consequences.

The interventions that show gains have something in common, and it is worth drawing out because it points to the design brief at the end of the chapter. Each of them repairs one of the breaks in the learning loop that Sterman described. The interactive simulator makes consequences visible and quick. The structured comparison makes the shared structure of different problems explicit, so that learners attend to it rather than to the surface. The courses teach the concept of accumulation directly, which unstructured experience never does. Unstructured experience, by contrast, leaves the loop as broken as it was: delays still separate action from result, outcomes are still confounded, and the learner's existing mental model still decides how the feedback is read. The difference between the pessimistic and optimistic evidence is, on this reading, a difference in how much the learning environment has been engineered to supply what the real system withholds.

The short-course evidence supports the claim that brief, targeted training improves stock-flow reasoning, and I'd grade that claim C. The pre/post studies are encouraging and their direction is consistent, but they lack randomised controls, their samples are highly selected, and a minority of participants still reason correlationally afterwards. They suggest that training helps; they do not yet provide an estimate of how much. Nor do any of them show that improved reasoning on inventory tasks leads to better decisions outside the classroom. Chapter 14 takes up that larger question, which is the weakest link in the evidence for systems-thinking education.

Does feeling that you understand mean that you do?

The evidence reviewed so far points to a specific danger, and it connects this chapter to the leverage illusion. Practice improves results, and improved results feel like understanding. A manager who has played a simulation

several times and scored better each time will reasonably believe that they have grasped how the system works. The Paich and Sterman results show that this belief can be false. The scores improved; the insight did not.

The same danger applies outside the laboratory, and there it is harder to detect. An experienced professional who has managed a system for years has accumulated a great deal of practice, and their results may be good for reasons that have little to do with understanding its structure: stable conditions, routines inherited from predecessors, or simple luck. The Moxnes and Brunstein results show that domain experience can coexist with poor reasoning about the dynamics of the domain. When conditions change, the practised routines may fail, and the confidence that experience has built may make the failure harder to recognise.

There is a simple test that anyone can apply to their own understanding of a system they manage, and it follows from the distinction between results and insight. Before making a change, write down what you expect to happen: which quantities will move, in which direction, by roughly how much and after what delay. Then compare the prediction with what happens. Good results with poor predictions suggest that your success owes more to conditions or routine than to understanding. Accurate predictions, especially for changes you have not made before, are better evidence that you have grasped the structure. The test is humbling, and it is useful for exactly that reason.

This is the leverage illusion in its experiential form. People come to feel that they have seen the structure of a system because they have spent time inside it, and they come to believe that they know where the levers are because they have pulled some of them and things turned out well. The evidence in this chapter suggests that neither belief follows from the experience alone. Structured experience can do better, but the gains it produces are gains in laboratory and classroom reasoning, and their connection to real decisions remains to be shown.

What should a practice environment give its users?

If unstructured experience teaches dynamics poorly and structured experience teaches it better, the practical question is how to design the structure. What follows is a design brief for a practice environment of the kind sometimes called a management flight simulator: a simulation in which people can make decisions about a dynamic system and see the consequences, without the costs of learning on the real thing. The brief draws on the evidence above, which supports it at the level of the C-grade training evidence: encouraging laboratory and classroom results, without controlled tests of transfer to real decisions.

The first question is whether to build one at all. A practice environment is worth the effort when three conditions hold together. The real system has long delays, so that learning from real outcomes would take years. Mistakes in the real system are costly or irreversible, so that experimenting on it is unacceptable. And the people who manage it make repeated decisions of the same broad kind, so that practice can build a transferable habit. Where these conditions do not hold, a simpler tool, such as the stock-flow sketch of Chapter 1 or the supply-line accounting of Chapter 3, is likely to deliver more for less.

Where they do hold, the environment should provide seven things:

1. Visible stocks. Show the accumulations explicitly, alongside the flows, throughout the exercise. The formats that Chapter 2 showed to support accurate reasoning belong in the simulator's displays as well as in reports.
2. Compressed time. Let users see, within a session, consequences that would take years to unfold in reality. Compression addresses the most important break in the learning loop that Sterman identified, the delay between action and result.
3. Feedback on the structure as well as the score. After each run, show users why the results came out as they did: which stocks rose or fell, which delays were at work, what their own decisions contributed. The Paich and Sterman evidence suggests that scores alone improve results without improving insight.

4. Room to experiment. Encourage users to try deliberately different strategies and compare them, including doing nothing. The Diehl and Sterman working paper found that people under pressure to control experimented very little, so the environment should make experimenting cheap and expected.
5. Comparison across cases. Give users more than one scenario with the same underlying structure and different surfaces, and ask them to identify what the cases share. The analogy evidence suggests that such comparisons reduce stock-flow errors, provided the shared structure is made explicit.
6. A test of understanding separate from performance. Before and after the exercise, ask users to predict how the system will respond to a change they have not practised. Improvement in prediction is better evidence of learning than improvement in score.
7. A debrief that connects the simulation to the real system. Discuss which features of the simulated system correspond to the real one, and which do not. A simulation that users cannot map onto their own situation teaches them to play the simulation.

Consider a hypothetical example. A regional water utility manages a reservoir system in which decisions about releases and restrictions affect storage levels over months, and droughts, when they come, reveal errors that were made long before. A practice environment for its operators would show reservoir storage as a stock with inflows from rainfall and outflows from consumption and release, compress a multi-year sequence of wet and dry periods into an afternoon, and let operators try different release rules, including very cautious ones and very aggressive ones. After each run it would show how storage evolved and why, and it would ask operators, before and after, to predict what would happen to storage under a rule they had not tried. The utility would then have a way of telling whether its operators' understanding had improved, not merely whether they had learned to play the simulation well.

A practice environment also carries a risk of its own, and it should be stated plainly. A simulator teaches the structure that its designers built into it. If that structure is wrong, because a delay is too short, a feedback loop is missing or a relationship has been guessed rather than estimated, then the environment will teach its users, efficiently and convincingly, to manage a system that does not exist. The more engaging the simulator, the more firmly the wrong lessons will stick. Anyone commissioning a practice environment should therefore ask how its structure was tested against the real system, a question that Chapter 15 examines in the context of much larger models, and should treat the debrief's discussion of where the simulation departs from reality as a central part of the exercise.

The last two items in the brief deserve emphasis, because they address the danger this chapter has identified. A practice environment that measures only performance will produce better performance on itself and a comfortable feeling of understanding. One that also measures understanding, and connects the simulation to the real system in the debrief, gives its users and its designers a way of finding out whether the feeling is justified.

Part II. Where the Promise Came From

The expectation that a clearly seen system will offer up a lever has a family tree, and this part climbs it using the founders' own texts. Each tradition met here struck its own bargain between rigorous results and broad ambitions, and the promise of control tended to grow fastest where the results underneath it were thinnest. One tradition, the soft school described in Chapter 9, declined to make the promise at all.

Chapter 6. The Cybernetic Moment: Feedback, Regulation and the First Promise of Control

Every thermostat makes a quiet philosophical claim. It behaves as if it wanted something, a room at a particular temperature, and it pursues that want by comparing what is with what should be and acting on the difference. For most of the history of science, talk of wanting and pursuing belonged to people and perhaps animals, and applying it to machinery was a figure of speech. In the 1940s a small group of mathematicians, engineers and physiologists argued that it was nothing of the kind. Purpose, they said, could be defined precisely, as a property of a certain kind of organisation, and that organisation was the same whether it was found in a nervous system, a steering engine or a heating plant. The name they gave to the study of such organisation was cybernetics, and the promise they made, or that others made on their behalf, is the earliest version of the promise this book is examining.

This chapter tells the story through four founding texts and the arguments that grew around them. It is intellectual history, and I want to be explicit at the outset about what that means for the book's grading. Nothing in this chapter is an empirical claim about whether an intervention works, and so nothing here receives an evidence grade. The cyberneticians produced definitions, theorems, thought experiments and a few ingenious devices. Their results are either mathematically true under stated conditions or they are metaphors, and the most important thing a practitioner can take from the period is the ability to tell those two apart. The leverage illusion, as the front matter defined it, is the belief that seeing a system's structure hands

you a reliable lever. Its first appearance comes when the mathematics of the regulator, which really does deliver levers inside engineering, is carried into firms and societies without the conditions that made it work.

Purpose without mystery

The first text is short. In 1943 Rosenblueth, Wiener and Bigelow published a paper in *Philosophy of Science* titled "Behavior, Purpose and Teleology" (Rosenblueth et al. 1943). Its argument was that purposeful behaviour can be defined through negative feedback: a system acts purposefully when some of its output is returned to it as information and used to reduce the gap between its present state and a goal. On that definition, a teleological description, one that explains behaviour by what it is aiming at, stops being unscientific. It becomes a description of a feedback arrangement, equally applicable to machines and to organisms.

It helps to see what the definition does and does not require, and a small illustration makes the point. Consider, as a hypothetical rather than a historical reconstruction, a device whose job is to keep a pointer aimed at a moving object. An open-loop device would compute where the object ought to be, point there, and never check. A closed-loop device measures the angle between pointer and object, turns the pointer in the direction that reduces the angle, measures again and repeats. The second device looks as if it is trying to follow the object. It corrects its errors, it recovers when bumped, and an observer who knew nothing of its wiring might well say that it wants to stay on target. Rosenblueth, Wiener and Bigelow's claim was that this description is accurate, and that there is nothing more to purpose, at the level of behaviour, than the loop that measures error and acts against it.

Two features of the definition matter for everything that follows. The first is that negative feedback is corrective: it counteracts deviation from a goal. A loop of this kind tends toward stability, and the whole of the early cybernetic programme was built on it. The second is that the goal sits outside the loop. The pointer does not choose which object to follow, and the thermostat does not decide what temperature is comfortable. Someone sets the goal, and the loop pursues it. In engineering this is so obvious that it hardly needs

saying. In a hospital or a ministry, where who sets the goal and whether it is the right one are often the hardest questions, it is the first condition to fall away when the language travels.

A common language for the animal and the machine

The second founding text is the book that gave the field its name. Wiener's *Cybernetics, or Control and Communication in the Animal and the Machine* appeared in 1948 and proposed that feedback, information and control could serve as a common language for machines and organisms (Wiener 1948). The subtitle states the ambition compactly. Wiener was reporting that thermostats and reflexes share a structure, and he was also proposing that a single vocabulary, built around the flow and use of information in loops, could unify the study of communication and control wherever it occurred.

That ambition was pursued in company. Between 1946 and 1953 a series of interdisciplinary meetings, the Macy conferences, brought together the people who gave cybernetics its early shape, and Heims's history of the group remains the standard account of how the field was formed in conversation among disciplines that rarely met (Heims 1991). The breadth of that conversation was the source of cybernetics' early energy and, as the historians show, of its later difficulties. A vocabulary that fits everything invites people to apply it everywhere, and the question of whether a given application is a result or a resemblance is easy to lose.

It is worth pausing on the difference between Wiener's two nouns, control and communication. Control is the part with mathematical teeth. Where a system's behaviour can be written down and its disturbances characterised, feedback control can be designed, analysed and proven to be stable or unstable under stated conditions. Communication, in the cybernetic sense, is what makes control possible: the loop needs information about the gap between goal and state, delivered in time to act on it. Both words were soon being used about economies, cities and families. Used carefully, they point at real questions: whether anyone in an organisation actually receives timely information about the outcomes they are supposed to be steering, for

example. Used loosely, they suggest that because an organisation can be described in the vocabulary of control, it can be controlled in the way a servomechanism can.

Richardson's intellectual history of feedback thinking is the best guide to how the vocabulary spread, and it makes a distinction that this book relies on (Richardson 1991). He traces two threads of feedback thought separately. One is the servomechanism thread, which runs from control engineering into system dynamics and is the subject of Chapter 8. The other is the cybernetic thread, which runs through Wiener and Ashby into biology, psychology, management cybernetics and the social sciences. The two threads share the idea of the loop, but they use it differently. The servomechanism thread cares about the accumulations and delays that shape how a loop behaves over time. The cybernetic thread cares more about information, variety and the logic of regulation. Much confusion in later systems writing comes from treating the two as one.

Ashby's homeostat and the law of requisite variety

The third founding contribution is a body of work by W. Ross Ashby rather than a single text, and it is the part of cybernetics that has aged best. In *Design for a Brain*, Ashby set out the idea of ultrastability and built a machine to embody it, the homeostat (Ashby 1952). His claim was that adaptive behaviour can be understood as a property of how feedback is organised. In Ashby's account, an ultrastable system has a first loop that regulates in the ordinary way and a second, slower mechanism that comes into play when certain essential variables are driven outside acceptable limits; the second mechanism changes the way the first loop is set up, and it keeps changing it until the system settles into a configuration that holds the essential variables within bounds. Adaptation, on this view, needs no designer inside the machine. It needs a structure that searches for stability when stability is lost.

The homeostat matters to this book as an anchor because it shows cybernetics at its most honest. Ashby built something, and the something did what his theory said it would do. The claim was modest in scope: a physical arrangement of feedback could exhibit behaviour that looked

adaptive. It did not claim that brains were homeostats, only that the principle was sufficient to produce adaptation in at least one concrete case. That is the pattern of a sound demonstration, and it is worth holding onto as the chapter moves toward uses of Ashby's ideas that had no device behind them.

Ashby's second great contribution was a textbook and a law. *An Introduction to Cybernetics* set out the field's concepts with unusual clarity and stated the law of requisite variety, which Ashby summarised in the principle that only variety can absorb variety (Ashby 1956). Two years later he gave the law a journal statement and drew out its implications for regulating complex systems (Ashby 1958). Variety, in Ashby's usage, is the number of distinguishable states something can be in. A regulator's job is to keep some outcome within an acceptable set despite disturbances. The law says, in its simplest form, that the variety of outcomes cannot be pushed below the variety of the disturbances divided by the variety of the regulator's possible responses. If the world can throw ten kinds of disturbance at a system and the regulator has only two responses, then some of the disturbances must get through to the outcome. No cleverness in choosing between the two responses can change that, because the limit is set by counting.

This is a genuine result, and its power comes from its generality. It does not depend on what the system is made of or what the regulator is trying to achieve. It applies to a thermostat that can only switch a heater on or off, which is why such a device cannot deal with a room that is too hot. It applies to a rulebook with a fixed menu of responses and to a call centre whose scripts cover some customer problems and not others. Its practical message is double. A regulator facing a varied world can either increase its own variety, by adding responses, discretion or local judgment, or it can reduce the variety that reaches it, by filtering, standardising or buffering the disturbances. There is no third way of holding outcomes steady. Chapter 16 shows how Stafford Beer built an entire theory of organisation, the Viable System Model, on this idea, and it is the version of cybernetics with the clearest line to practical use.

Every good regulator must be a model

The fourth founding text pushed the logic of regulation one step further. In 1970 Conant and Ashby published a paper whose title states its conclusion: "Every good regulator of a system must be a model of that system" (Conant and Ashby 1970). They offered a theorem showing that, under broad conditions, any regulator that is both maximally successful and maximally simple must be isomorphic with the system it regulates. Put less formally, the best and leanest possible regulator has to contain, in some form, a representation of what it is regulating.

The theorem is often quoted as if it said that managers need models, and that is roughly the spirit in which it entered systems writing. The spirit is not wrong, but the letter is narrower. The result concerns regulators that are optimal on two criteria at once, and it establishes a structural correspondence under the stated conditions. It says nothing about how accurate a real, imperfect regulator's model needs to be, how the model should be built, or whether an organisation that invests in modelling will regulate better than one that does not. The fair summary is that the theorem is a formal result, and that how far it reaches into management practice has been argued rather than shown. I would add only that the argument has usually been made by people who wanted to build models, which is not a reason to reject it but is a reason to notice that it has not been tested.

Taken together, requisite variety and the good regulator theorem form the strongest theoretical case cybernetics ever made for the value of understanding a system before trying to steer it. A regulator must command at least as much variety as the world throws at it, and the leanest successful regulator carries a model of what it regulates. Both statements are true in their formal settings. Both became, in popular form, arguments that anyone who understands a system's structure well enough can control it. That popular form is the leverage illusion in its first recognisable shape, and the distance between the theorem and the slogan is where most of the trouble lies.

Amplifiers, observers and the scattering of a movement

Within twenty years the founders' frame was being widened from several directions at once. The first widening came from Maruyama, who argued in 1963 that cybernetics had paid too much attention to deviation-counteracting feedback, the corrective kind that stabilises, and too little to deviation-amplifying processes (Maruyama 1963). He called the study of these processes the second cybernetics. In a deviation-amplifying process, a small initial difference is fed back in a way that makes it larger: a slight advantage attracts resources that increase the advantage, or a small decline triggers responses that deepen the decline. Maruyama's point was that such mutual causal processes explain a great deal of growth, differentiation and divergence in biological and social systems, and that a theory built only on stabilising loops cannot account for them.

The widening was important for two reasons. It completed the basic grammar of feedback: loops that counteract and loops that amplify, which Chapter 10 later formalises as balancing and reinforcing loops with their own notation. It also changed the character of the promise. A world of stabilising loops suggests control by setting goals and letting regulators pursue them. A world that also contains amplifying loops suggests that small, well-placed pushes might grow into large effects. Maruyama was describing mechanisms, not recommending interventions, but the idea that a small deviation can be amplified into a large change is recognisably an ancestor of Meadows's phrase about a small shift producing big changes in everything, which Chapter 12 examines.

The second widening turned cybernetics back on itself. Von Foerster's essays set out what came to be called second-order cybernetics, in which the observer is treated as part of the system being observed, and questions of self-reference and the construction of knowledge move to the centre (von Foerster 2003). If the first cybernetics studied systems that regulate, the second studied the systems that describe regulation, including the scientists themselves. Maturana and Varela's account of autopoiesis added a biology of self-producing organisation, in which a living system continuously produces the components that constitute it (Maturana and Varela 1980). Autopoiesis

was later carried into social theory, where its fit is disputed. Bateson, meanwhile, applied cybernetic ideas to mind, levels of learning and ecology in a collection of essays that proved influential in family therapy and in second-order cybernetics itself (Bateson 1972).

These developments are intellectually rich, and they deserve better than a summary. For the argument of this book, however, their significance is specific. Second-order cybernetics undercut the assumption, built into the first, that the regulator stands outside the thing it regulates and sets goals from a position of knowledge. Once the observer is part of the system, the picture of a controller reading structure and choosing where to push becomes much harder to sustain. In that sense the cybernetic tradition developed its own internal critique of the leverage illusion before the illusion had become fashionable, and Chapter 9 shows how the soft systems tradition drew similar conclusions by a different route.

The historians' accounts explain why these ideas did not hold together as a single science. Kline's history of the American side describes the rise of cybernetics and information theory and their subsequent fragmentation, and, as its subtitle announces, asks why we came to call ours the information age (Kline 2015). Pickering's history of British cybernetics, covering Walter, Ashby, Bateson, Laing, Beer and Pask, reads the British work as a performative science, more interested in building things that adapt than in representing the world accurately (Pickering 2010). Heims's account of the Macy group shows a movement whose participants were united by a vocabulary more than by a method (Heims 1991). Read together, the three histories describe a field that scattered into control engineering, computing, cognitive science, management cybernetics, family therapy and social theory, carrying its vocabulary into each and leaving its mathematics mostly behind in the first two.

From mathematics to metaphor

The scattering is where the mathematical and the metaphorical versions of cybernetics parted company, and it is worth being precise about what the mathematics requires. A feedback controller in engineering works reliably

when several conditions hold. The controlled variable can be measured accurately and quickly. The goal is known and fixed, or changes in a known way. The controller's actions affect the variable through a relationship that is stable and can be characterised. The delays between action and effect are short relative to the speed of the disturbances, or at least are known and can be compensated for. The controller stands outside the system, so that its own behaviour is not reshaped by the thing it is controlling. And the disturbances are limited in kind, so that the controller's repertoire of responses can match them.

When these conditions hold, feedback control provides real levers. An engineer can adjust a gain, add a sensor or shorten a delay and predict the result with confidence. That is leverage in a strict and useful sense, and nothing in this book casts doubt on it. The trouble starts when the vocabulary is transferred to settings where most of the conditions fail. In a large organisation, the variables that matter are often hard to measure and slow to report. Goals are contested, and the people nominally in charge of setting them disagree. The relationship between action and outcome shifts as people adapt to the action. Delays are long and poorly understood; Chapter 3 showed how badly people manage even a short, known supply-line delay. The would-be controller is inside the system, subject to its incentives and its information. And the variety of disturbances is vast.

A hypothetical example shows how quickly the conditions fall away. Suppose a regional education department decides to treat school attendance the way an engineer treats room temperature. It sets a target rate, measures attendance monthly, and when a school falls below target it sends an improvement adviser, much as a thermostat switches on a heater. On paper this is a negative feedback loop, and a diagram of it would look reassuringly like a control system. In practice the measurement arrives weeks after the absences it records, so the adviser responds to last term's problem. The schools learn which absences are counted and adjust how they record them, so the measured variable drifts away from the real one. The reasons children stay away vary from illness to transport to family crisis, and a single adviser with a standard toolkit has far fewer responses than there are causes. And the target itself is contested, because some staff think attendance is a

symptom rather than the thing to be managed. Every condition that made the thermostat reliable has quietly failed, yet the vocabulary of loops and targets survives intact, and that survival is what makes the arrangement look controllable when it is not.

None of this means the cybernetic concepts are useless outside engineering. Requisite variety, in particular, can be applied to organisations without pretending that they are servomechanisms, because it is a counting argument and does not depend on the conditions for stable linear control. What goes wrong is the inference from description to control. An organisation can be described as a set of feedback loops, and the description can be illuminating. The inference that because it can be so described, a sufficiently knowledgeable manager can steer it the way an engineer steers a plant, is the step that has no support, and it is the step that the rest of this book follows as it recurs in later traditions.

I think the cyberneticians themselves were more aware of this gap than many of their popularisers. Ashby's homeostat demonstrated a principle and offered no blueprint for management. Conant and Ashby's theorem was carefully conditional. Maruyama was describing mechanisms, and the second-order cyberneticians were actively sceptical of external control. The promise that knowledge of structure yields control was assembled from their results by others, and often by stripping out the conditions. That is why the right response to cybernetics is to keep its results and their conditions together, and to regard any application that drops the conditions as untested until someone tests it.

Auditing a regulator's variety

The most portable result from this period is the law of requisite variety, and it can be turned into a short audit of any arrangement meant to keep something under control. The audit is a structured way of counting whether a regulator's repertoire is as varied as the trouble coming at it, and it applies equally to a rota, a triage protocol, a customer-service script, a regulatory regime or a quality-assurance process. It rests on theory, not on outcome studies, and I would present it to colleagues as a way of organising judgment

rather than as a validated diagnostic. No study has tested whether organisations that perform such audits regulate better than those that do not. The audit runs in five steps.

1. **Name the essential variable and its acceptable range.** State what the arrangement exists to hold steady, such as waiting times below a threshold or defects below a rate, and who decided the range. If no one can say, the audit stops here, because a regulator without a goal cannot be assessed.
2. **List the disturbances.** Write down the distinct kinds of event that push the essential variable out of range. Group them only where they genuinely call for the same response. This list is the disturbance variety.
3. **List the responses.** Write down the distinct actions the regulator can actually take, including those that require someone's discretion, and note who can take each one and how quickly. This list is the response variety.
4. **Match the two lists.** For each disturbance, identify a response that neutralises it. Disturbances with no matching response are the ones that will reach the outcome, whatever the regulator does.
5. **Choose a remedy for each gap.** Either add response variety, by creating a new response or giving someone the discretion to improvise one, or reduce disturbance variety, by filtering, buffering or standardising the inputs so that fewer kinds of event reach the regulator. Record which you chose and why.

A worked example shows how the steps play out. Consider a hypothetical hospital outpatient clinic whose managers want to keep the average wait for a first appointment under a set number of weeks. Suppose that when the managers list the disturbances, they find six distinct kinds: seasonal surges in referrals, sudden staff sickness, a new referral pathway from primary care, equipment failure in one diagnostic room, a backlog left by a previous cancellation, and referrals that arrive incomplete and must be returned. Suppose that when they list the responses, they find only three that the clinic can take on its own authority: offering overtime, reshuffling the appointment book, and asking referrers to resend incomplete forms.

Matching the lists shows at once that three of the six disturbances have no response the clinic controls. Staff sickness, equipment failure and the new pathway will reach the waiting list regardless of how skilfully the existing three responses are used.

The remedies then divide along Ashby's two routes. The clinic could add response variety, for example by agreeing in advance with a neighbouring clinic that it can borrow a diagnostic slot when its own equipment fails, or by giving a senior nurse the authority to reorganise sessions when staff are absent. Or it could reduce disturbance variety, for example by working with primary care so that the new pathway sends referrals in a standard format that fits the existing booking process. The example is invented, and the numbers in it are there only to make the counting visible. The point is that the audit turns the vague question of whether a system is under control into a specific question that can be answered by listing and matching.

Two cautions keep the audit honest. First, it measures capacity to regulate, not the quality of regulation. A clinic with enough responses on paper may still use them badly, too slowly or in the wrong order, and Chapter 3's evidence on delays suggests that timing is where many real regulators fail. Second, the audit assumes that the essential variable is the right one. A clinic that holds average waiting times steady by shortening appointments may have perfect variety for the wrong goal. That second caution is where cybernetics hands over to the traditions of the next three chapters, which ask in turn how complex systems are organised, how their behaviour unfolds over time, and whose goals the regulators are serving.

Chapter 7. General Systems and the Architecture of Complexity

Is there a science of wholes? The question sounds grand, and for a generation after the Second World War some serious scientists thought the answer was yes. If the same patterns of organisation turned up in cells, organisms, firms and societies, then perhaps a single body of theory could describe them all, and a person trained in that theory could move between disciplines, recognising the same structures under different names. A different group of thinkers, working at the same time and sometimes in the same institutions, gave a more guarded answer. They agreed that complex systems had properties worth studying in their own right, but they argued that the most useful thing to understand about complexity was how it could be broken down, and how far the pieces could be studied separately.

This chapter sets those two visions side by side. The first is general system theory, associated above all with von Bertalanffy and Boulding, which sought laws that would hold across disciplines. The second runs from Weaver's distinction between kinds of complexity through Simon's account of hierarchy, and it produced a quieter, more testable set of claims. Like the previous chapter, this one deals in theory. None of the ideas here has been tested as a claim about interventions, so none receives an evidence grade, and I have tried to present each on its own terms before asking what it implies for the leverage illusion.

The first vision: laws that hold everywhere

Von Bertalanffy's early English-language statement of general system theory, published in 1950, argued that open systems in biology and in society share formally similar laws, and that a general science of wholes was therefore

needed alongside the specialised sciences (Bertalanffy 1950). An open system, in his sense, is one that exchanges matter and energy with its environment and maintains itself through that exchange, as a living organism does. His collected statement of the theory, published in 1968, gathered the ideas that became its signature: open systems, equifinality (the capacity of an open system to reach the same final state from different starting conditions and by different paths), and isomorphisms, the formal similarities between laws in different fields (Bertalanffy 1968).

The idea of isomorphism is the heart of the first vision, and it is easy to see its appeal. If the equation describing the growth of a population has the same form as the equation describing the spread of a rumour or the adoption of a technology, then a result proved in one domain might transfer to the others. A general theory would supply the shared forms, and the specialised sciences would supply the content. For practitioners, the promise was that a trained systems thinker could recognise a familiar pattern in an unfamiliar setting and know something useful about it without starting from scratch.

Boulding's 1956 essay, "General Systems Theory: The Skeleton of Science", gave the programme its most memorable image (Boulding 1956). He presented general systems theory as a skeleton on which the flesh of the particular sciences could hang, and he proposed a nine-level hierarchy of increasing system complexity. What is often forgotten is how carefully Boulding hedged. He explicitly disclaimed any ambition to produce a theory of everything, which puts him closer to the cautious second vision of this chapter than his image of a skeleton might suggest.

The general systems movement also had social aims, and Hammond's history of the founders of the Society for General Systems Research treats those aims as central to understanding the movement, alongside its scientific ones (Hammond 2003). Its title calls the project a science of synthesis. Social aims of that kind explain some of the movement's energy, and they also explain why its claims sometimes ran ahead of its results. A programme that hopes to unify the sciences and improve the world has strong reasons to see isomorphisms everywhere.

Seen from the vantage point of this book, the first vision carries a particular risk. If the same laws hold everywhere, then someone who has learned them in one place can apply them in another with confidence. That is exactly the move by which a pattern recognised in a diagram becomes a lever pulled in a real organisation. The risk lies less in the theory itself, which its founders often stated with care, than in how it was used: as a licence to transfer conclusions between domains without checking whether the similarity of form carried a similarity of behaviour.

A hypothetical case shows how the transfer can mislead. Suppose an analyst notices that enrolments in a new public training scheme follow the same S-shaped curve as the spread of an infection: slow at first, then rapid, then levelling off. The formal similarity is real, and the same family of equations can describe both curves. It would be tempting to borrow the epidemiologist's toolkit wholesale and conclude, say, that enrolment will level off once a fixed share of the eligible population has joined, just as an outbreak burns out once enough people are immune. But the mechanisms may differ completely. The levelling off in enrolments might come from a cap on places, from a funding cycle or from word of mouth turning negative, and each of these responds to a different intervention. The shared curve says nothing about which mechanism is at work. An isomorphism of form is a prompt to look for a shared mechanism, and it becomes knowledge only when the mechanism has been checked in the new setting.

The second vision: organised complexity and nearly separable parts

Weaver's 1948 essay "Science and Complexity" drew a distinction that shaped much of what followed (Weaver 1948). He separated three kinds of problem. Problems of simplicity involve a few variables, and classical physics had mastered them. Problems of disorganised complexity involve vast numbers of variables behaving more or less at random, and statistics and probability had learned to handle them by averaging. Between the two lay problems of organised complexity, involving a moderate number of variables that are interrelated in an organised way, so that neither the

methods for simple problems nor the methods for random aggregates would do. Weaver argued that problems of this third kind, which included much of biology and social life, were the great challenge for science.

The distinction is useful because it names a middle ground precisely. It explains why neither the reductionist habit of isolating a few variables nor the statistical habit of averaging over many is enough for a firm, an ecosystem or a hospital. It also explains why such systems resist the kind of prediction that physics achieved for planets and statistics for gas molecules. The practitioner's version of the point is that an organisation is too interconnected for a single-variable fix and too structured for averages alone to describe it.

Weaver's middle ground is also where Part I of this book found people struggling. The stock-flow and beer-game tasks of Chapters 1 to 3 are tiny examples of organised complexity: a handful of variables, linked by accumulation and delay, whose behaviour neither intuition about single causes nor intuition about averages predicts well. That is one reason the laboratory findings matter for the history told here. The founders of this second vision identified organised complexity as the hard case for science, and the experimental evidence later showed that it is a hard case for ordinary judgment too.

Anderson's 1972 essay "More Is Different" gave the physicist's statement of the same insight (Anderson 1972). Anderson argued that each level of complexity has properties of its own, so that accepting reductionism, the view that everything is made of the same basic constituents obeying the same laws, does not imply that wholes can be reconstructed from knowledge of their parts. The properties that appear at a higher level are real, and they have to be studied at that level. This is what later writers call emergence, and Anderson's version is valuable because it comes from inside physics, the discipline most often assumed to have reduced everything to its parts.

Simon took the idea of organised complexity and asked a more specific question: what structure do complex systems actually have, and why? His answer, set out in "The Architecture of Complexity" in 1962 and later reprinted in *The Sciences of the Artificial*, was hierarchy and near-decompos-

ability (Simon 1962; Simon 1969). A hierarchic system is composed of subsystems, which are themselves composed of subsystems, and so on down. A nearly decomposable system is one in which interactions within subsystems are much stronger and faster than interactions between them. In such a system, each subsystem can be understood, in the short run, largely on its own, and the interactions between subsystems matter mainly in aggregate and over longer periods.

Hora, Tempus and the advantage of stable parts

Simon made the argument vivid with a parable about two watchmakers, Hora and Tempus (Simon 1962). Both make fine watches of about 1,000 parts, both are highly regarded, and both are constantly interrupted by customers telephoning their workshops. Tempus builds each watch in a single sequence, so that when he puts a partly assembled watch down to answer the phone it falls to pieces and he must begin again. Hora designs his watches so that he can put together subassemblies of about ten parts each, combine ten of those into larger subassemblies, and combine ten of the larger ones into the whole watch, so that an interruption costs him only the subassembly in progress. In Simon's telling, Hora prospers while Tempus grows poorer and finally loses his shop, and a short calculation shows why: if there is one chance in a hundred of an interruption as each part is added, Tempus takes on average about four thousand times as long as Hora to finish a watch.

Simon drew two lessons from the parable, and both matter here. The first concerns evolution. Complex systems are much more likely to arise, whether by natural selection or by human design, if they are built from stable intermediate forms, and this explains why so many of the complex systems we observe are hierarchic. The second concerns understanding. Because near-decomposable systems can be studied subsystem by subsystem, they are comprehensible in a way that fully interconnected systems would not be. Hierarchy is what makes complexity tractable.

I find this the most important idea in the chapter for a practitioner, and it is worth saying why it bears so directly on the leverage illusion. Popular systems thinking often insists that everything is connected to everything else and that any attempt to manage a part in isolation is doomed. Simon's account says something more precise and more useful. Everything may be connected, but the connections vary enormously in strength and speed, and in many real systems the strong, fast connections cluster within subsystems while the weak, slow ones run between them. Where that is so, managing piece by piece is often a sensible strategy, and the parts can be improved without redesigning the whole. The places where piece-by-piece management fails are the places where near-decomposability breaks down: where strong interactions cross the boundaries between units, or where the slow interactions between subsystems accumulate into effects that dominate the long run.

This is a quieter claim than universal systems laws, and it is better founded, because it can be checked. For a given organisation, one can ask where the strong interactions are and whether they respect the formal boundaries between departments. A manufacturer's product design team and its production plant, for example, may sit in different divisions with different managers, but if every design change forces the plant to retool and every production constraint sends drawings back to the designers, then the interaction between them is strong and fast, and managing them separately will fail. The same firm's payroll office may interact with both only weakly and slowly, and managing it on its own may be entirely reasonable. The example is illustrative, but the question it models is general: where does near-decomposability hold, and where does it not?

Where small things do matter: increasing returns

The second vision also contains a formal account of how small events can have large effects, and it needs to be kept separate from the looser promise that small shifts produce big changes everywhere. Arthur's 1989 model of competing technologies showed that under increasing returns, where each adoption makes a technology more attractive to the next adopter, chance events early in a competition can decide which technology wins and keep it

in place, even when a rival would have served better (Arthur 1989). The outcome is path-dependent: it depends on the order in which early, essentially accidental events happened, and once lock-in occurs it is hard to reverse.

Arthur's result is a genuine and narrow sense in which small shifts matter. It specifies the conditions, increasing returns and an early period in which the competition is still open, under which a small push can tip a system toward one outcome. It also specifies the cost of the same mechanism: once the system is locked in, a push of the same size achieves almost nothing. Leverage, in Arthur's model, is concentrated in a window of time and depends on a particular kind of feedback, the reinforcing kind that Chapter 6 traced to Maruyama. Chapter 12 returns to this as one of the few formally grounded senses of leverage available to the field.

The broader literature on complex adaptive systems placed such mechanisms within a general picture. Holland's *Hidden Order* described complex adaptive systems in terms of agents, building blocks, tags and internal models, showing how adaptation at the level of individual agents can build complexity at the level of the whole (Holland 1995). Mitchell's guided tour of complexity science is notable for its candour: it presents the field's achievements clearly while admitting that its central terms, including complexity itself, still lack agreed definitions (Mitchell 2009). That admission is a useful corrective. A field that has not settled what complexity is cannot yet offer general laws for managing it, and the more careful complexity scientists have said so.

From science of systems to philosophy of management

The last movement in this lineage turned the systems idea from a scientific programme into a philosophy of management and planning. Churchman's *The Systems Approach* made the case accessibly, emphasising the practice of sweeping in considerations that a narrower analysis would leave out and the ethics of deciding where to draw a system's boundary (Churchman 1968). An often-quoted formulation of Churchman's holds that the systems approach begins at the moment one sees the world through the eyes of another

person, which makes perspective-taking, rather than technical modelling, the entry point. The emphasis on boundaries as ethical choices is the seed of the boundary critique that Chapter 9 develops.

Ackoff took a parallel path. His 1971 paper attempted to organise systems concepts themselves into a system, defining states, environments, behaviour, purposeful systems and organisations in relation to one another (Ackoff 1971). By 1979 he had turned sharply against his own discipline. "The Future of Operational Research is Past" argued that academic operational research had drifted from practice, lost its interdisciplinarity and failed to adapt to what Ackoff called messes, the tangled problem situations that practitioners actually face (Ackoff 1979). The paper is often read as a turning point toward interactive and soft systems approaches, and it prepares the ground for the next two chapters.

Surveys of the whole lineage, such as Jackson's account of the systems approaches and Ramage and Shipp's collection of about thirty systems thinkers, show how widely these traditions diverged once they entered management (Jackson 2003; Ramage and Shipp 2020). They are useful secondary maps. They also show that the first vision, of general laws, left a durable vocabulary and few durable results, while the second, of hierarchy and organised complexity, left fewer slogans and more ideas that can actually be checked against a particular case.

Sketching boundaries and hierarchy

The practical residue of this chapter is a sketch that any practitioner can make before deciding whether a problem needs a whole-system approach at all. Its purpose is to find where the system's boundary should be drawn and which parts can safely be treated as nearly independent. It draws on Simon's near-decomposability and on Churchman's view that boundaries are choices

with consequences. Like the requisite-variety audit in Chapter 6, it rests on theory rather than on outcome evidence, and it should be treated as a disciplined way to organise judgment. The sketch has four steps.

1. **State the behaviour you care about.** Write down the outcome that is causing concern, as a pattern over time if possible, such as waiting times that rise every winter or a product line whose quality erodes after each reorganisation.
2. **Draw a first boundary generously.** Include every unit, group or process that plausibly affects the behaviour. Then note, for each one you have left out, why you left it out. Churchman's point is that exclusions are decisions, and the reasons should be visible to others.
3. **Map the strength and speed of interactions.** For each pair of units inside the boundary, mark whether they interact strongly or weakly, and quickly or slowly. A simple grid will do. Strong, fast interactions show where units must be managed together. Weak, slow interactions show where units can be managed separately, at least in the short run.
4. **Test the formal structure against the grid.** Compare the pattern of strong interactions with the organisation's formal divisions. Where strong interactions cross a divisional boundary, piece-by-piece management is likely to fail, and those crossings are the places to look first. Where strong interactions stay within divisions, local improvement is probably a reasonable strategy, and a whole-system intervention may add cost without adding much.

Consider a hypothetical regional ambulance service to see the sketch at work. Its managers are worried that response times lengthen every winter. A generous first boundary includes the ambulance fleet, the control room, the hospitals' emergency departments, the hospitals' discharge teams and the community care services that take patients after discharge. Mapping interactions shows that the fleet and the control room interact strongly and quickly, which is expected and already managed together. It also shows a strong interaction that crosses organisational lines: ambulances wait outside emergency departments when those departments are full, and the departments are full because discharges from the wards are delayed by shortages in

community care. That chain crosses three organisations, and it is where piece-by-piece management of the ambulance service alone is unlikely to succeed. The same sketch shows that the service's vehicle maintenance interacts only weakly with the rest and can be improved on its own. The example is invented to show the logic, and a real sketch would need data on waiting and handover times to confirm which interactions are strong.

The sketch has a modest aim. It does not identify a lever, and it does not claim that the strong crossing interaction is the right place to intervene. What it does is tell a practitioner where the system can reasonably be decomposed and where it cannot, which is the question Simon's work equips us to ask and the question that general systems slogans tend to skip.

Chapter 8. Forrester's Wager: System Dynamics and the Idea of the Hidden Lever

A method, like a person, can be judged by what it promised when it was young and what it still believes when it is old. System dynamics made its first promise in 1958, when Forrester announced industrial dynamics to readers of the *Harvard Business Review* as a major breakthrough for decision makers (Forrester 1958). Chapter 3 followed one thread from that article, the production-distribution simulation that became the beer game and the long line of experiments built on it. This chapter follows the method itself. Its founding claim was that the troubles of a firm, its swings in production and inventory, could be produced by the firm's own structure of orders, delays and decision rules, without any help from the outside world. That was new to most managers, and it was true often enough to matter.

That claim is the seed of everything in this chapter. It grew first into a set of ambitious books that applied the method to cities and to the world. Along the way it acquired the idea that is central to this book's argument, the idea that complex systems behave counterintuitively, so that the lever that would improve them is hidden from ordinary judgment and intuition tends to push it the wrong way. The chapter follows the method's first books and their reception, and it has a specific aim. It tries to separate a durable scientific idea, which deserves real credit, from an overreach that the critics identified at the time and that the field has since partly acknowledged.

Factories as feedback systems

Forrester's first book, *Industrial Dynamics*, set out the method in full in 1961 (Forrester 1961). Its central move was to treat a firm as an information-feedback system. Decisions about production, hiring, ordering and pricing are made on the basis of information about the state of the firm, such as inventory levels, backlogs and workforce. Those decisions change the state of the firm, which changes the information, which changes the next round of decisions. Seen this way, a firm's behaviour over time, its growth, its cycles, its crises, is generated by the loops that connect its decisions to its states. The production-distribution amplification of the 1958 article appears again as a worked demonstration of how such loops can turn a steady world into a turbulent firm.

Three features of the approach made it different from what came before, and all three survive in the field today. The first was the emphasis on accumulations. The state of a firm is held in stocks, of inventory, workers, orders or cash, which change only through flows, and the delays in those flows shape behaviour. This is the same structure Chapter 1 introduced with the bathtub, and its importance to system dynamics is one reason the field's researchers later took such interest in whether people can reason about it. The second was the treatment of decisions as policies. Forrester modelled managers' decisions as rules that respond to available information, which made it possible to ask how changing a rule would change the firm's behaviour. The third was the endogenous point of view: the insistence that the explanation for a system's behaviour should be sought mainly in its internal feedback structure, and only secondarily in outside shocks.

A hypothetical firm shows what the endogenous point of view adds. Suppose a manufacturer sees its output and workforce swing up and down over a cycle of two or three years, and its managers blame the swings on fluctuating demand. An industrial dynamics analyst would ask a different question: could the firm's own policies produce the cycle even if demand were steady? Suppose the firm sets production to close the gap between its actual inventory and a target, and hires workers when production needs rise, but hiring and training take several months. A small, one-off rise in

orders draws inventory down. The firm raises its production target and starts hiring, but the new workers arrive only after the delay, by which time the firm has fallen further behind and hired still more. When the new workers are all in place, output overshoots, inventory builds up above target, and the firm cuts back, again with a delay, until it overshoots in the other direction. The cycle is generated by the interaction of an inventory rule, a hiring rule and a delay, and it would persist with perfectly steady demand. The example is invented, but its logic is the one Forrester's 1958 simulation displayed, and it shows why the endogenous point of view changes what managers look for. If the firm's own rules make the cycle, then forecasting demand more accurately will not remove it, and changing the rules might.

Richardson's intellectual history of feedback thought places this work in what he calls the servomechanism thread, the line of feedback thinking that came from control engineering and cared about how accumulations and delays shape a loop's behaviour over time (Richardson 1991). That distinguishes it from the cybernetic thread of Chapter 6, with its focus on information and variety. The two threads share the loop, but system dynamics took from engineering a habit of simulating behaviour over time that cybernetics, in its social applications, mostly did not.

The treatment of decisions as rules drew a charge that system dynamics ignored the theory of decision making. Morecroft answered it in 1983 by showing how system dynamics decision rules could be read as portraits of bounded rationality in the tradition of the Carnegie school: rules of thumb that use limited, local information, as real managers do (Morecroft 1983). The answer matters because it connects the method to the behavioural evidence. When Sterman later tested a behavioural simulation model by having people make the modelled capital investment decisions themselves, he found that a proposed heuristic explained their choices well (Sterman 1987). The link between simulated decision rules and observed human decisions, established in experiments like that one and the beer-game work of Chapter 3, is part of what makes the durable idea durable.

Urban Dynamics and the housing result

The method's second major book moved from the firm to the city. *Urban Dynamics*, published in 1969, presented a simulation model of an urban area and drew from it a conclusion that was bound to attract attention: that building low-income housing worsens urban decline (Forrester 1969). A policy intended to help the city's poorest residents would, according to the model, make the city as a whole worse off over time.

It is worth being clear about the form of this argument, because it became the template for the counterintuitive claim. The model linked the city's stocks and activities through feedback loops. A policy that looks beneficial in its immediate effects triggers responses elsewhere in the structure, and those responses, playing out over years, overwhelm the immediate benefit. The conclusion is presented as something the model reveals and intuition misses, and it carries an implied prescription: the intuitively attractive policy should be abandoned, and something that looks less attractive should be done instead.

This is the anchor case of the chapter because the housing result was the first time system dynamics made a strong, specific and politically charged policy prescription, and because its reception showed where such prescriptions are vulnerable. A firm's managers can test an industrial dynamics model against their own records of orders and inventory, and they share, more or less, a view of what the firm is trying to achieve. A city is different. Its data are patchier, its causal relationships are harder to establish, and its residents disagree about what a better city would be. Each of those differences became the basis for a critique.

Counterintuitive by design

In 1971 Forrester generalised the lesson in an article titled "Counterintuitive Behavior of Social Systems", published in *Theory and Decision* (Forrester 1971a). Its argument was that human intuition about complex feedback systems is systematically wrong, and that as a result policies often have the opposite of their intended long-run effect. On this view the intuitive policy, aimed at the obvious symptom, triggers feedback that eventually undoes it.

This article is where the leverage idea took its modern form, and it helps to spell out the logic in steps. If complex systems are counterintuitive, then the places where intervention would work are not where intuition looks for them. The lever is hidden. Worse, because intuition is systematically wrong, people who do find the right lever are likely to push it in the wrong direction. Only a formal model of the system's feedback structure can reveal where the lever is and which way it should be pushed. The practical conclusion follows naturally: policy should be designed with models, because unaided judgment cannot be trusted with complex systems.

Donella Meadows later passed on one of Forrester's favourite stories, which she presented as part of the lore that everyone trained by him had absorbed: time after time he had located a leverage point in a company, only to find that people there were already paying close attention to it and pushing it in the wrong direction (Meadows 1999). She told it near the start of her essay, before her list, as the first example of the backward intuition that Forrester called counterintuitive behaviour, and followed it with the world model's verdict on growth and the Urban Dynamics verdict on subsidised housing. Chapter 12 quotes and analyses that passage, so I note it here only because it shows how directly the counterintuitive thesis fed into the most widely read text on leverage.

The same year brought *World Dynamics*, published by Wright-Allen Press, which applied the method to the entire globe (Forrester 1971b). Its World2 model, prepared for the Club of Rome, was the precursor of the World3 model behind *The Limits to Growth*. Its best-known critic, Nordhaus, attacked it in 1973 under the title "World Dynamics: Measurement Without Data", arguing that its relationships had not been estimated from evidence and that its conclusions depended on untested assumptions about matters such as technology and substitution (Nordhaus 1973). Forrester and colleagues replied the following year (Forrester, Low and Mass 1974). Chapter 15 takes up that exchange and its sequel in full, and grades Nordhaus's critique there.

It is important to see what is right in the counterintuitive thesis before turning to its problems. The laboratory evidence of Part I shows that people do misjudge accumulation, delay and feedback, reliably and across populations, and that the misjudgements are of exactly the kind Forrester's argument needs: people treat stocks as if they tracked their inflows and ignore the supply line of orders already placed. If intuition about feedback systems is that unreliable in tasks as small as a bathtub or a four-stage supply chain, it would be surprising if it were reliable in cities. The thesis also has many compelling examples, gathered over decades, of policies that were resisted or reversed by the systems they targeted (Sterman 2006; Sterman 2001). Policy resistance, the tendency of a system to defeat interventions through feedback that the intervener did not anticipate, is a real phenomenon that deserves a name.

What the thesis lacks is a base rate. Nobody has established how often intuitive policies backfire, as opposed to how often they work as intended, or how often model-designed policies do better. The examples are compelling, but examples are selected for being compelling. Every field that collects cases of its own insight accumulates a gallery of successes, and a gallery of policies that backfired is no different. What it cannot show is the denominator: the many intuitive policies that did roughly what their designers intended, which no one writes up as counterintuitive because they were not. A thesis that says social systems are counterintuitive in general needs evidence about social systems in general, and that evidence does not exist. I'd grade the claim C: the case examples are rich and the laboratory work on misperception makes the mechanism plausible, but there is no systematic base rate, and the method's own most famous policy conclusion, the Urban Dynamics housing result, was strongly challenged by critics who showed how much it depended on the modeller's choices.

Three critics and a requiem

The critics came quickly. The most penetrating, in my judgment, was Kadanoff, whose 1971 paper "From Simulation Model to Public Policy" examined how Urban Dynamics moved from its equations to its recommendations (Kadanoff 1971). His argument was that the model's counterin-

tuitive results followed directly from Forrester's normative scheme, the criterion used to judge whether one outcome for the city was better than another. Change the scheme, Kadanoff showed, and the conclusions reverse.

This is the single most important critique for the argument of this book, and it is worth dwelling on. The counterintuitive thesis presents the model as revealing a hidden truth about the system: that low-income housing makes the city worse off. Kadanoff's analysis showed that the model's definition of worse off was an input, supplied by its author. A city with fewer poor residents and more industry might score well on one measure of urban health and badly on another, depending on whose welfare the measure counts. The lever the model identified, and the direction in which it said to push, were therefore partly consequences of a value judgment, and a different value judgment pointed the lever the other way. The model's structure could be sound in every causal detail and its prescription could still be contestable, because a prescription needs a goal and the goal came from outside the structure. This is the point Chapter 6 made about thermostats, whose goals are set by someone, arriving in a policy setting where the goal is precisely what people disagree about.

A deliberately simple hypothetical makes the mechanism of reversal concrete. Suppose a model of a city projects two futures over twenty years. Under the first policy, the city ends up with more jobs, a higher average income and fewer low-income residents, because some have left for places where housing is cheaper. Under the second, it ends up with fewer jobs, a lower average income and more low-income residents, most of whom are better housed than before. The model's projections are the same whichever way they are judged. If the criterion is the city's average income or its tax base, the first policy wins. If the criterion is the welfare of the people who were poor at the start, wherever they end up living, the second may win. Nothing in the causal structure of the model settles which criterion to use. The choice is a political and ethical one, and a model that embeds one criterion without saying so will present a value judgment as a discovery. Kadanoff's contribution was to show that Urban Dynamics had done something of this kind, and that the counterintuitive recommendation depended on it.

The second critique came in 1972. Gray, Pessel and Varaiya re-ran Urban Dynamics, testing it under rent control, changes to its parameters and an exogenous labour force, and criticised both its economics and its validation (Gray et al. 1972). Their re-runs asked a simple question: how much do the model's conclusions depend on particular assumptions that could reasonably have been made differently? Their answer was that they depended on them a good deal, and that the model had not been checked against urban data in a way that would justify confidence in the assumptions.

The third was broader still. In 1973 Lee published "Requiem for Large-Scale Models" in the *Journal of the American Institute of Planners* (Lee 1973). His subject was large-scale urban modelling as an enterprise, and his verdict was that none of the goals of large urban models had been achieved and that planning needed different methods. The requiem is not a critique of Urban Dynamics in particular, and I cite it for that reason: it shows that the doubts raised about one model were part of a wider reckoning with the idea that big simulations could steer urban policy.

Taken together, the three critiques identified the overreach precisely. None of them denied that cities contain feedback loops or that intuitive policies can backfire. What they denied was that a model built largely from its author's judgment, with limited calibration against data and a built-in normative scheme, could support confident policy prescriptions. Forty years later, Featherston and Doolan, surveying the published objections to the method from a sympathetic standpoint, judged that some of them still held, notably those concerning the role of historical data, reductionism, and plurality and hierarchy (Featherston and Doolan 2012). That is a notable admission, and it was made by people who wanted the method to succeed. The concern about historical data is the one that bears most directly on Urban Dynamics: a model whose parameters are not estimated from data can reproduce plausible behaviour without giving strong grounds for believing that its specific prescriptions are right.

Reflection, schooling and a profession

Forrester's later writing shows a founder reflecting on what the method had achieved, and it contains a shift in emphasis that is easy to miss. In "Lessons from System Dynamics Modeling", published in 1987, he argued that modelling pays off when it yields generalisations that change people's mental models, and that it fails when it stops at describing a model (Forrester 1987). The counterintuitive response of systems to policy appears there as one of the lessons. The emphasis has moved from models as devices for designing policy toward models as devices for changing how people think. That is a more defensible position, and it anticipates the use of simulation for learning that Chapter 5 examined.

The same shift runs through the reflections of the next generation. Sterman's essay on becoming a systems scientist, whose title reminds readers that all models are wrong, argues for humility about models, for testing them hard and for being frank about their limits (Sterman 2002). Read beside the confident prescriptions of *Urban Dynamics*, it shows how far the field's self-understanding had moved in three decades. The later position does not abandon the claim that models can inform policy. It changes the terms on which they do so, asking the modeller to treat a model's recommendation as something to be examined and tested rather than as a finding to be implemented.

In 1993 Forrester made the case for system dynamics as the organising framework for pre-college education (Forrester 1993). The proposal expressed his conviction that the counterintuitive nature of systems was the most important thing a citizen could learn, and that the earlier it was learned the better. Whether systems thinking can be taught, and whether teaching it changes anything that matters, is the subject of Chapter 14. In 2007 he published a personal view of the field's first fifty years, describing how it began and recounting its cornerstone projects (Forrester 2007). It is an essential primary source and a partisan one, as founders' retrospectives usually are.

Meanwhile the field professionalised. Sterman's *Business Dynamics*, published in 2000, became the standard graduate text, covering the modelling process, stocks and flows, delays, the beer game, validation and the misperceptions of feedback that his own experiments had documented (Sterman 2000). A shorter statement in the *California Management Review* explained policy resistance and the use of management flight simulators as tools for learning (Sterman 2001). Barlas proposed a formal scheme for judging the validity of such models, in which a model's structure is examined before its behaviour is compared with data, and he observed that validation had received surprisingly little explicit attention in the field until then (Barlas 1996). Chapter 15 returns to the harder question of what validating a model of an open social system can mean.

Two reflective papers from inside the field show how it came to understand its own foundations. Lane took on the charge that system dynamics is a hard or deterministic approach, distinguishing four readings of that charge and arguing that the method offers conditional predictions, of the form "if this policy, then this behaviour", rather than prophecy (Lane 2000). Richardson revisited Forrester's founding threads and argued that the field's core consists of four kinds of thinking: dynamic, stock-and-flow, operational and endogenous (Richardson 2011). Both papers read as efforts to state what the method is, now that its founding ambitions have been tempered by experience.

The durable idea and the overreach

The biography of the method now allows a verdict, and it has two parts that need to be kept apart.

The durable idea is that endogenous feedback structure explains behaviour. A system's patterns over time, its oscillations, overshoots, growth and collapse, are often generated by its own internal loops of accumulation, delay and decision, and looking for those loops is usually more productive than blaming outside shocks. This idea has held up remarkably well. The production-distribution amplification of 1958 became the beer game, and the beer game became one of the most replicated findings in behavioural

operations, as Chapter 3 showed. The stock-flow structure at the heart of the method turned out to describe precisely the kind of problem that people find hardest, as Chapters 1 and 2 showed. The decision rules of system dynamics models have been tested against human decisions in the laboratory and found to fit. Forrester deserves full credit for this idea. It changed how a generation of researchers and managers thought about the causes of their troubles, and much of the best evidence in this book exists because he insisted on it.

The overreach is policy prescription from uncalibrated models. The counterintuitive thesis moved from the true observation that intuition about feedback is unreliable to the much stronger claim that a model built by an expert could reliably find the hidden lever and the direction to push it. That stronger claim ran into three problems that the critics identified and the field has partly acknowledged. Models built largely from judgment, with limited estimation from data, can generate plausible behaviour without giving good grounds for trusting their specific prescriptions. The direction in which a model says to push depends on the goal used to judge outcomes, and in public policy that goal is contested, as Kadanoff showed. And the claim that social systems are counterintuitive in general has no base rate behind it, only well-chosen examples.

The leverage illusion, in the form this book is tracing, is precisely the fusion of these two parts. It takes the durable insight that structure drives behaviour and treats it as though it implied that seeing the structure reveals the lever. Forrester's later emphasis on changing mental models rather than prescribing policies suggests that he came to see some of the gap himself. The practitioner's task is to keep the two parts separate: to take the endogenous point of view seriously as a way of looking for causes, and to treat any specific lever that a model identifies as a hypothesis that needs its own evidence, including evidence about whose goal it serves.

Two of the criticisms that Featherston and Doolan judged still pertinent point beyond this chapter. The charge of reductionism concerns what a model leaves out: every model draws a boundary, and the variables outside it are treated as given even when they respond to the policy being tested. The

charge about plurality and hierarchy concerns whose view of the system the model encodes. A system dynamics model usually represents one causal story, agreed by the modellers and perhaps their clients, and it can struggle to hold several stakeholders' incompatible accounts of the same situation. That second charge is where the soft systems tradition of Chapter 9 begins. Its founders argued that in many real problem situations there is no single agreed system to model, and that a method which assumes one will produce a lever that some of the people involved do not recognise as a lever at all.

It would be unfair to leave the impression that system dynamics simply ignored these points. The field responded in several ways: by building models with client groups rather than for them, a practice Chapter 11 examines; by shifting emphasis from prescription toward learning, as Forrester's own later writing did; and by developing standards for testing and documentation. Whether those responses closed the gap is a question for Part III. What the history of the first books establishes is the shape of the gap itself. On one side stands a method with a genuinely powerful idea about where behaviour comes from. On the other stands a habit, formed in its most ambitious early applications, of treating the output of a single expert model as a reliable guide to where and which way to push. The chapters that follow test the second against the evidence, and they keep returning to the first as the part worth keeping.

Five steps of a modern modelling project

System dynamics as it is taught today, following Sterman's textbook and his shorter statement of the method, proceeds through five broad steps (Sterman 2000; Sterman 2001). They are the practice standard of the field, and they embody many of the lessons that the Urban Dynamics debate taught. Knowing them lets a practitioner commission, read or challenge a model intelligently, whether or not they build models themselves. There is

no evidence grade to attach to the steps as such; they describe good practice, and whether projects that follow them produce better decisions has not been tested.

1. **Problem articulation.** Define the problem as a pattern of behaviour over time, the reference mode, and state the time horizon over which it matters. Choose the boundary: which variables will be explained by the model and which will be taken as given.
2. **Dynamic hypothesis.** Propose an explanation of the reference mode in terms of the system's feedback structure, its stocks, flows, delays and decision rules. The hypothesis should be endogenous where possible, explaining the behaviour through internal structure.
3. **Formulation.** Turn the hypothesis into a simulation model, specifying equations, parameters and initial conditions, and noting which are estimated from data and which are judgments.
4. **Testing.** Check whether the model reproduces the reference mode, whether it still gives plausible results when inputs are pushed to extreme values, and whether its structure corresponds to what is known about the real system. Barlas's advice to test structure before behaviour belongs here.
5. **Policy design and evaluation.** Use the model to explore how changes in decision rules or structure would change behaviour, test how robust the results are to uncertainty in the assumptions, and state the goal against which policies are being judged.

A hypothetical example shows how the steps would handle a problem, and where the Urban Dynamics lessons come in. Imagine a regional health service that is losing nurses faster than it can train them. In problem articulation, the team defines the reference mode as a rising vacancy rate over the past eight years and sets a ten-year horizon. It draws the boundary to include the nursing workforce, training places and workload, and leaves national pay policy outside as a given. The dynamic hypothesis proposes a reinforcing loop: vacancies raise the workload on remaining nurses, higher

workload raises the rate at which nurses leave, and more leavers mean more vacancies. It adds a slow balancing loop in which rising vacancies eventually prompt more training places, but only after a long delay.

In formulation, the team builds stocks for trained nurses, nurses in training and vacancies, with flows for hiring, graduation and leaving. It estimates the leaving rate from the service's own records and marks the relationship between workload and leaving as a judgment, because the records do not show it directly. In testing, it checks that the model reproduces the eight years of rising vacancies and that it behaves sensibly in extreme cases, for instance that vacancies do not become negative if hiring is set very high. In policy design, it compares expanding training places with reducing workload through temporary staff, and finds, suppose, that the second breaks the reinforcing loop faster because training places take years to produce nurses.

The Urban Dynamics debate adds three questions that the team should answer before anyone acts on that finding. First, how much does the conclusion depend on the judged relationship between workload and leaving, which was not estimated from data? A sensitivity test should show whether the ranking of the two policies survives plausible alternative values. Second, whose goal is the model optimising? A service judged on its vacancy rate might prefer temporary staff, while one judged on continuity of care might not, and that is the kind of normative choice Kadanoff showed can reverse a conclusion. Third, what evidence would show that the chosen policy worked, and when would it arrive? A model's recommendation is a hypothesis about the real system, and the project is incomplete until that hypothesis has been tested against what happens next. The example is invented, but each of its three questions comes directly from the critics who first examined what system dynamics claimed.

Chapter 9. The Soft Turn: When There Is No Single System to See

Try a short exercise before reading on. Pick an organisation you know well, a hospital, a school, a council or a firm, and write one sentence that states what it is for. Then write the same sentence three more times, as it might be written by a long-serving frontline worker, by the person who funds the organisation, and by someone who uses its services and is unhappy with them. If your four sentences agree, the organisation is unusual. More often they describe different purposes, and each purpose implies a different system: different boundaries, different measures of success, different ideas of what counts as a problem. A model built on the first sentence would identify levers that the authors of the other three might regard as irrelevant, or even as part of the problem.

That exercise captures the argument of the tradition this chapter describes. Through the 1970s and 1980s a group of researchers in operational research and systems practice argued that the hard systems methods of the previous two decades rested on an assumption that often failed. The assumption was that there is a system out there, with a definite structure and an agreed purpose, which the analyst's job is to model and improve. In many of the problems that mattered most, the soft tradition replied, there was no agreed purpose and therefore no single system to model. Different stakeholders saw different systems, and the analyst's job was to help them understand one another's views well enough to agree on action.

The soft tradition partly rejected the leverage illusion before anyone named it. If different stakeholders see different systems, there is no single structure from which a lever can be read. The chapter is organised as a dialogue between the soft tradition and the hard tradition it criticised, and it ends

with the critical-systems writers who tried to reconcile them. It also records, as honestly as the sources allow, what evidence exists that the soft methods work.

Problems that refuse to hold still

The clearest statement of the problem came from outside operational research. In 1973 Rittel and Webber published "Dilemmas in a General Theory of Planning" in *Policy Sciences*, and introduced the term wicked problems (Rittel and Webber 1973). A wicked problem, in their account, has no definitive formulation: the way the problem is stated already shapes what solutions are considered. It has no stopping rule, so there is no point at which it is clearly solved. Its solutions can only be better or worse, never true or false, and different parties will judge them differently. And every attempt at a solution is consequential, because it changes the situation and cannot be undone for the purposes of a clean experiment. From these properties Rittel and Webber drew the conclusion that planning cannot follow the engineering paradigm, in which a problem is defined, analysed and solved.

The argument was aimed at planning, but its implications for systems methods were direct. The engineering paradigm is the one that Chapter 6 traced to feedback control and Chapter 8 to system dynamics' early policy models. It assumes that the goal is known, that the system can be characterised, and that the analyst can test solutions against the goal. A wicked problem violates all three assumptions. Its goal is contested, its formulation depends on who is formulating it, and its solutions can only be tried once, in the real world, with real consequences. Kadanoff's demonstration, described in Chapter 8, that the conclusions of Urban Dynamics reversed under a different normative scheme is a concrete instance of the same point: the model's lever pointed one way or the other depending on whose idea of a better city was built into it.

Ackoff reached a similar conclusion from inside operational research. Chapter 7 described his 1979 complaint that the discipline had lost touch with practice and with the messy situations practitioners meet, and that

complaint belongs in this story too (Ackoff 1979). By the end of the 1970s, then, two lines of argument had converged on the same diagnosis. Many real problems come without a settled shape or a shared goal, and methods that presume both will fail in ways their users may not notice.

Checkland's reversal: systems as ways of knowing

The most developed response came from Checkland and his colleagues at Lancaster, and its founding statement was *Systems Thinking, Systems Practice*, published in 1981 (Checkland 1981). Checkland drew the distinction between hard and soft systems thinking that gave the tradition its name, and he made a philosophical move that is easy to state and hard to absorb. In hard systems thinking, systems are taken to exist in the world, and the analyst's models are attempts to describe them. In soft systems thinking, systems are epistemological devices: ways of organising our thinking about a messy situation, which do not claim to describe how the world is actually structured. The word "system" moves from the world to the analyst's head.

The consequence for method was radical. If a system is a way of thinking, then there can be several, each corresponding to a different view of what the situation is about. Soft systems methodology (SSM), as Checkland developed it, works by building several models of purposeful activity, each based on a different view, and using them to structure a debate among the people involved about what changes would be both desirable and feasible. The models serve as tools for learning and make no claim to predict, and the method ends in an accommodation among people with different views.

Checkland's own thirty-year retrospective describes how the method developed from action research at Lancaster into a mature learning cycle, built around three main devices (Checkland 2000). The first is the rich picture, a drawing of the problem situation that tries to capture its structures, processes, relationships and conflicts without forcing them into a model too early. The second is the root definition, a concise statement of a relevant system of purposeful activity, checked against a mnemonic, CATWOE, that prompts the analyst to identify the elements any such system must have. The third is the purposeful activity model, a set of

logically connected activities that a system matching the root definition would need to carry out, which is then compared with what actually happens in the real situation to structure the debate about change. The retrospective is valuable as a record of how the method's originator understood it, and it should be read as that and not as an independent evaluation.

Checkland and Scholes's *Soft Systems Methodology in Action* presented the method through extended cases from organisations including the National Health Service, ICL and Shell, written by the developers themselves (Checkland and Scholes 1990). Checkland and Poulter's *Learning for Action* later gave a concise statement of the method for practitioners (Checkland and Poulter 2006). The cases matter because SSM was developed through practice, and its claims rest largely on its developers' accounts of what happened when they used it. That is a strength, in that the method was shaped by real problems, and a limitation, in that the accounts come from the people with the greatest investment in the method's success.

A soft systems study in outline

The mechanics of SSM are easiest to grasp through an example. Checkland's published cases are the natural source, but their details could not be checked for this book, and rather than reconstruct them from memory I use a hypothetical study, constructed to show how the devices fit together.

Consider a hypothetical community mental health service whose waiting times have risen and whose staff, managers, referring doctors and service users all describe the situation differently. A hard systems study would begin by defining the service's objective, probably reducing waiting times, and would model the flow of referrals through assessment and treatment. A soft systems study begins more slowly. The analyst spends time with each group and draws a rich picture: referring doctors frustrated by long waits and unclear criteria; clinicians overwhelmed by caseloads and worried that pressure to discharge patients quickly is causing relapses; managers answerable to a funder who monitors waiting times; service users who value continuity with a single clinician and experience the service as a sequence of

handovers. The picture shows conflicts that a flow model would not capture, such as the tension between the funder's measure and the clinicians' concern for relapse.

From the rich picture, the analyst identifies several relevant systems, one for each distinct account of what the service exists to do. One root definition might describe a system for assessing and treating referred patients within target times, so that the funder's standards are met. Another might describe a system for providing continuous therapeutic relationships to people with long-term mental health problems, so that relapse is reduced. A third might describe a system for supporting referring doctors in managing patients who do not need specialist care. Each is checked against CATWOE: who benefits or suffers from the system, who carries out its activities, what transformation it achieves, what worldview makes that transformation meaningful, who could stop it, and what constraints in its environment it takes as given.

The analyst then builds a purposeful activity model for each root definition and compares each with what actually happens. The comparison is the heart of the method. It shows, for instance, that the service does many of the activities in the first model and few of those in the second, and it gives staff, managers and service users a structured way to discuss whether that balance is the one they want. The outcome might be a set of changes that none of the groups would have proposed alone: a separate pathway for long-term patients that protects continuity, for example, alongside a triage arrangement with referring doctors that reduces inappropriate referrals. What the study does not produce is a single lever, because it never assumed a single system to which a lever could belong.

The hard tradition answers back

The soft tradition's critique of hard methods did not go unanswered, and the answers have force. The most careful came from system dynamics. Lane's defence of the method against the charge of being hard, summarised in Chapter 8, turned on the difference between a conditional forecast and a prophecy (Lane 2000). For the dialogue of this chapter the more pointed

part of his reply is practical: system dynamics had long been used for learning as well as for prediction, and its modellers often worked alongside their clients, which blurs the line the soft tradition drew between the two camps.

Lane and Oliva went further and made a theoretical case for combining system dynamics with SSM (Lane and Oliva 1998). Their argument was that each method covers the other's blind spots. SSM is good at surfacing different views and building accommodation, but it has little to say about how a situation will behave over time once a change has been made. System dynamics is good at exploring behaviour over time, but it tends to assume a single agreed model. A synthesis would use SSM to structure the debate about purposes and system dynamics to test the dynamic consequences of the changes being debated. This was, and remains, an argument rather than an empirical result, but it is a sensible one.

A broader family of methods grew up alongside SSM, and Rosenhead's 1989 collection *Rational Analysis for a Problematic World* gave it a name: problem structuring methods (Rosenhead 1989). The collection and its successor, edited with Mingers in 2001, brought together SSM, strategic options development and analysis (SODA), strategic choice, robustness analysis and related approaches, all designed to help groups structure problems that lack an agreed formulation (Rosenhead and Mingers 2001). Their shared premise was that analysis in problematic situations should help people decide what the problem is, and should not assume that the problem is already known. Chapter 11 examines group model building, a system dynamics method that shares this participatory character and is sometimes classed with the problem structuring methods.

How widely these methods were taken up varied sharply by place. Mingers's 2011 assessment described Soft OR as established in the United Kingdom and parts of Europe but marginal in the United States (Mingers 2011). The geography is worth noting, because it means that much of the literature on soft methods comes from a particular academic and practice community, and that practitioners elsewhere may never have encountered the methods at all.

The hard tradition has one further reply, and I think it is the strongest. Agreement among stakeholders is not evidence that the agreed change will work. Part I of this book showed that intelligent, educated people misjudge accumulation and delay in simple tasks, and they do so in the same direction, so a room full of them can reach a confident consensus that is dynamically wrong. The hypothetical mental health study described above might produce an accommodation that every group endorses, such as adding clinical staff, without anyone noticing that the backlog will take months to respond and that the new staff will be judged a failure before their effect appears. Nothing in the logic of SSM protects against this, because the method's test of a proposed change is whether the participants find it desirable and feasible, and participants' sense of feasibility is exactly what the stock-flow evidence says is unreliable. This is the practical content of Lane and Oliva's argument for combining the traditions. The soft methods are right that there is often no single agreed system to model, and the hard methods are right that, once a change has been agreed, its consequences unfold through accumulations and delays that no amount of consensus will alter.

Whose system? Boundary critique and critical pluralism

The critical systems writers took the soft tradition's insight one step further. If systems are ways of thinking and different people think differently, then drawing a system's boundary is a value judgment as well as a technical choice, and value judgments can be made well or badly, fairly or unfairly. Churchman had already argued, as Chapter 7 described, that boundary setting carries ethical weight (Churchman 1968). Ulrich turned the argument into a method.

Ulrich's *Critical Heuristics of Social Planning* laid the philosophical foundation (Ulrich 1983), and his 1987 summary set out its practical core: twelve boundary questions, each asked in two modes, "is" and "ought" (Ulrich 1987). The questions probe the value judgments built into a design, about such matters as whose purposes it serves and who has a say in it, and they are meant to give a voice to those affected by a design but not involved in making it. Asking each question in the "is" mode reveals the value

judgments currently built into a design. Asking it in the "ought" mode invites the participants to say what those judgments should be. The comparison surfaces the choices that a design embodies and gives a voice to people who would otherwise be left out of the conversation. Critical Systems Heuristics, as the approach is known, is designed to make boundary judgments visible and open to challenge, especially by those who bear their consequences.

Midgley developed these ideas into what he called systemic intervention, organised around boundary critique and the use of several methods together (Midgley 2000). His argument was that no single systems method is adequate to every situation, and that a good intervention begins by examining boundaries and then chooses methods to fit what the examination reveals. The same argument for methodological pluralism underlies Flood and Jackson's Total Systems Intervention, launched in 1991 as a meta-method for choosing among systems approaches according to the character of the problem situation (Flood and Jackson 1991). The book could be checked only through its catalogue record, so I describe its purpose and no more.

Jackson's later books gave the critical-systems position its most comprehensive form. *Systems Thinking: Creative Holism for Managers* surveyed the range of systems approaches, including hard systems methods, system dynamics, the Viable System Model, SSM and Critical Systems Heuristics, and argued for critical pluralism: using different approaches for different purposes, with awareness of what each assumes (Jackson 2003). *Critical Systems Thinking and the Management of Complexity* organised the approaches by the kinds of complexity each addresses (Jackson 2019). Flood, for his part, applied critical-systems and complexity ideas to a reappraisal of Senge's learning organisation, a subject Chapter 13 takes up (Flood 1999). Reynolds and Holwell's edited guide to five systems approaches for managing change offers a practical entry to several of these methods side by side (Reynolds and Holwell 2010).

The critical-systems synthesis is the place where the dialogue of this chapter comes to rest, and its position on the leverage illusion is clear. A lever exists only relative to a boundary and a purpose. Different boundaries and purposes produce different levers, and choosing among them is a matter of judgment and, often, of power. A method that identifies a lever without making its boundary judgments explicit has made a value choice silently. That is an important correction to the hard tradition, and it anticipates many of the problems Part IV finds when big levers are pulled.

The soft tradition in the dock

The soft and critical traditions deserve credit for their insight, and they also deserve the same scrutiny this book applies to every other part of the field. What evidence is there that these methods work?

The most direct evidence is a survey. Mingers and Taylor surveyed about 300 people who had been exposed to SSM, and nearly half replied; most of those who replied had used the method and judged it successful (Mingers and Taylor 1992). The finding is genuine, and it is evidence of something: people who use SSM tend to value it. It is also exactly the kind of evidence that is weakest for judging effectiveness. The respondents were self-selected, since those who had found the method useful were probably more likely to reply. Their judgments of success were their own, not measured against any outcome. And there was no comparison group of people who had used another method or none.

Mingers and Rosenhead's review of published applications of problem structuring methods reached a similar conclusion about the wider family (Mingers and Rosenhead 2004). The evidence it found was overwhelmingly practitioner case reports: accounts, usually by the facilitators, of how a method was applied in a particular organisation and what the participants made of it. Case reports can be rich and informative, and they are how practice-based methods are usually documented. They cannot show whether the method did better than an alternative would have done, because there is no alternative in the account.

To its credit, the field has put its own methods on trial. Ackermann's 2012 paper, titled "Problem Structuring Methods 'in the Dock'", weighed the case for and against Soft OR and acknowledged its thin evaluation record (Ackermann 2012). That willingness to examine the evidence openly is more than some other parts of the systems field have managed, and it deserves recognition. The comparative gap it identifies is also visible in the neighbouring literature on group model building: the most recent review of that method reports that no study has set it against any of its sister methods, a point Chapter 11 develops (Scott et al. 2016).

I'd grade the evidence C for use and self-reported success and D for comparative effectiveness. The C reflects consistent reports that practitioners value these methods and find them workable, from a survey and a large body of case accounts, but with self-selected respondents and no independent outcome measures. The D reflects the complete absence of designs comparing soft methods with alternatives on any outcome, proximal or distal. The soft tradition's philosophical critique of the leverage illusion stands on its own merits, since it is an argument and not an empirical claim. Its claim to have a better way of intervening is, so far, a hypothesis that its own practitioners have acknowledged needs testing.

It is worth sketching what a better test would look like, because the difficulty is real and it is not unique to soft methods. The outcome of an SSM study is an accommodation among people, and the value of that accommodation depends on what happens afterwards. A serious evaluation would compare organisations or groups that used SSM with similar ones that used another structured method, or their usual approach, on the same kind of problem. It would measure proximal outcomes, such as participants' understanding of one another's views and their commitment to the agreed changes, using measures other than the participants' own satisfaction. It would also follow up on distal outcomes: whether the changes were implemented, whether they lasted, and whether the problem situation improved by the participants' own earlier standards. Such designs are expensive and hard to arrange, and the problems they would study are, by definition, ones where success is contested. Those difficulties explain the thin record, but they do not remove the need for it. The group model

building literature of Chapter 11 shows that at least some comparative and controlled designs are possible for participatory methods, and the soft tradition has not yet produced its equivalent.

Worksheets for purpose and boundary

The soft and critical traditions leave two tools that any practitioner can use, whether or not they adopt SSM or Critical Systems Heuristics in full. The first is a CATWOE worksheet for writing and comparing root definitions. The second is a set of boundary questions asked in "is" and "ought" modes. Both are designed to make visible the purposes and boundaries that any intervention assumes. Their evidence base is the one described above, C for practitioners' reports of their value and D for evidence that they improve outcomes compared with alternatives, and I would use them on that understanding.

A CATWOE worksheet. For a problem situation, write at least two root definitions, each expressing a different view of what the relevant system is for. For each one, fill in six entries, following Checkland's mnemonic as set out in his accounts of the method (Checkland 1981; Checkland and Poulter 2006):

- **Customers:** who benefits from, or is harmed by, what the system does?
- **Actors:** who carries out the system's activities?
- **Transformation:** what input does the system change into what output?
- **Worldview (*Weltanschauung*):** what belief about the world makes this transformation meaningful and worth doing?
- **Owners:** who could stop the system or change it fundamentally?
- **Environmental constraints:** what does the system take as given from outside, such as budgets, laws or other organisations?

Then compare the root definitions entry by entry. The comparison usually reveals where the definitions share actors and constraints but differ in worldview and in whom they count as customers, and those differences are what the stakeholders need to discuss before any change is designed.

Boundary questions in two modes. The questions below are my adaptation, written in the spirit of Ulrich's twelve; practitioners who want the original wording should take it from Ulrich's own statements (Ulrich 1983; Ulrich 1987). Ask each question twice, first as "what is the case?" and then as "what ought to be the case?", and record both answers.

1. Whose interests does the intervention serve, and whose should it serve?
2. What is its purpose, and what should its purpose be?
3. By what measure will success be judged, and by what measure should it be judged?
4. Who decides on the intervention's design and resources, and who should decide?
5. What expertise does it rely on, and what expertise should it rely on?
6. What guarantees its success, if anything, and what would a genuine guarantee require?
7. Who is affected by the intervention without being involved in designing it, and how should their interests be represented?
8. What worldview does the intervention take for granted, and what alternative worldviews are being set aside?

A hypothetical example shows how the two tools combine. Suppose a university wants to reduce the number of first-year students who drop out, and has commissioned an analysis that identifies stricter entry requirements as the most effective lever. The CATWOE worksheet produces two root definitions. In one, the relevant system selects students likely to complete a degree, its customers are the university and its funders, and its worldview holds that completion rates measure institutional quality. In the other, the system helps students from many backgrounds to succeed once admitted, its customers are the students themselves, and its worldview holds that a university exists to widen opportunity. The boundary questions, asked in both modes, show that the analysis measures success by the completion rate alone, that no current or former students were consulted, and that its boundary leaves out the applicants who would be turned away. None of this shows that stricter entry requirements are wrong. It shows that the lever the

analysis identified belongs to the first root definition and not the second, and that choosing it is a decision about purpose and about whose interests count, which is where the soft tradition always said the real choice lay.

Part III. The Toolkit on Trial

The methods examined in the next five chapters all claim, in one way or another, to show where a system can be moved. Each is taught first on its own terms and then set against what has been published about it, beginning with diagrams and shared models, passing through Meadows's essay on where to push, and ending with packaged patterns, the organisations meant to use them, and the question of whether the underlying skill can be taught at all.

Chapter 10. Loops on Paper: What Causal Loop Diagrams Show and Hide

Nearly every systems method in this book eventually produces a picture. Some produce equations as well, and a few produce simulations, but the picture is what gets shown to the board, pinned to the wall of the workshop room and reproduced in the report. The most common picture by far is the causal loop diagram: words joined by curved arrows, each arrow carrying a small plus or minus sign, with the arrows closing into loops that are labelled as reinforcing or balancing. It is the first tool most people learn when they meet systems thinking, and for many it is the only one they ever use.

This chapter does two things with it, in order. It first teaches the notation properly, because a critique of a tool is worth little if the reader has never been shown how the tool is meant to work. Earlier chapters described feedback in plain words; this is where the book introduces the formal notation. It then turns to the warnings that have come from inside the system dynamics field itself, the best known of them about 40 years old, about what these diagrams conceal. The argument of the second half is narrow but important for the rest of Part III. A causal loop diagram is a good way to record a hypothesis about structure, and a poor way to infer behaviour from it. Because claims about levers are claims about behaviour, the step from a tidy loop to a confident recommendation is exactly the step the diagram cannot support on its own.

Drawing an arrow and giving it a sign

A causal loop diagram has only three kinds of element. There are variables, written as short noun phrases naming quantities that can rise or fall. There are arrows, each claiming that one variable influences another. And there

are signs on the arrows, called link polarities, which say in which direction the influence runs. Loops appear when a chain of arrows returns to its starting point. Everything else in the diagram, including the loop labels and any delay marks, is derived from these three.

The polarity of a link is where most of the meaning sits, and where most of the trouble starts, so it is worth defining carefully. A positive link from A to B means that if A increases, B will be higher than it would otherwise have been, and if A decreases, B will be lower than it would otherwise have been. A negative link means the reverse: an increase in A leaves B lower than it would otherwise have been, and a decrease leaves it higher. The phrase "than it would otherwise have been" looks like pedantry. It is the most important clause in the definition, and a later section shows what goes wrong when readers drop it, as most of them do.

To make the notation concrete, consider a hypothetical outpatient clinic. The numbers and relationships that follow are invented for illustration and do not describe any real service. The clinic keeps a waiting list of patients who have been referred but not yet seen. When the list grows, the pressure felt by staff rises, so the link from the waiting list to staff pressure is positive. When pressure rises, managers schedule extra clinic sessions, another positive link. More sessions mean more patients seen per week, positive again. And more patients seen means a shorter waiting list than there would otherwise have been, which is a negative link. Written in a line, the chain reads: waiting list, plus, staff pressure, plus, extra sessions, plus, patients seen, minus, back to the waiting list.

Variable names deserve more care than beginners usually give them. A good name is a quantity with a direction, such as "staff pressure" or "patients seen per week", and not an action such as "increase staffing" or a mood such as "crisis". The test is whether the phrase "more of" or "less of" can sensibly be put in front of it. A diagram built from actions and moods cannot be given consistent polarities, because it is unclear what an increase in "crisis" would mean or how it would be measured. Units help as well. Writing "patients" beside the waiting list and "patients per week" beside patients seen costs a few seconds and prevents a confusion that the chapter returns to below.

Loops, their signs and the delay mark

Once the arrows close into a loop, the loop itself has a polarity, and there is a simple rule for finding it. Count the negative links around the loop. If the count is zero or even, the loop is reinforcing: a change anywhere in it comes back around to push the same variable further in the same direction. If the count is odd, the loop is balancing: a change comes back around to oppose itself. The clinic chain above has one negative link, so it is a balancing loop. A longer waiting list produces responses that shorten it. Practitioners mark it with a B, or with a small circular arrow and a minus sign, drawn inside the loop.

The two loop types correspond to the two kinds of feedback that Chapter 6 traced through the cybernetic literature. Balancing loops are the corrective, goal-seeking loops of the early cyberneticians. Reinforcing loops are the deviation-amplifying processes that Maruyama (1963) argued the field had neglected: loops that produce growth, decline, vicious circles and virtuous ones. In a diagram they look alike, and only the count of signs tells them apart.

The hypothetical clinic almost certainly has a reinforcing loop as well. Suppose that sustained staff pressure also means more overtime, that overtime produces fatigue, that fatigue produces more errors, and that errors create rework, such as repeat appointments and corrected letters, which adds to the workload and so to the pressure. Every link in that chain is positive. With no negative links, the loop is reinforcing, and it runs in whichever direction it is pushed. Under rising pressure it makes things worse; if pressure eases, the same loop helps things improve faster. Marked R, it sits beside the balancing loop and shares the variable “staff pressure” with it.

That shared variable is where a diagram starts to become interesting and where it starts to become dangerous. The clinic now has two loops pulling in different directions from the same point. Which one wins? The balancing loop says pressure brings the list down; the reinforcing loop says pressure makes the work harder and so keeps the list up. The honest answer is that the diagram cannot say. Which loop dominates depends on magnitudes the

diagram does not contain: how many extra sessions a given level of pressure buys, how quickly fatigue sets in, how much rework an error creates, and how all of these change over time. Loop dominance can also shift, so that the balancing loop controls the system for months and then the reinforcing loop takes over once fatigue passes some point. A causal loop diagram draws the candidates for dominance. It does not rank them.

The last standard element is the delay. Many practitioners mark a delay by drawing a pair of short parallel lines across the arrow where it occurs. In the clinic, extra sessions do not appear the week pressure rises. Rotas have to be changed and rooms booked, so the link from pressure to sessions carries a delay mark. Chapter 3 showed what delays in a balancing loop do to people who manage it: they overshoot, because the correction they ordered has not yet arrived when they decide whether to order more. A balancing loop with a significant delay tends to oscillate rather than settle smoothly. The delay mark is a warning that this can happen, but, like the loop label, it does not say how large the swing will be or whether it will damp out.

Richmond (1993), in the paper that gave the field much of its modern skills vocabulary, listed closed-loop thinking among the core skills of a systems thinker, alongside dynamic thinking and operational thinking. The causal loop diagram is the natural tool of closed-loop thinking. It trains the eye to look for the ways in which an effect comes back to influence its cause, which is a real and valuable habit given the evidence in Part I that people tend to reason in open chains. What it does less well is operational thinking, if that term is taken, as I read it, to mean attention to how a system physically works: what accumulates, what flows and what the flows are made of. That weakness is the subject of the rest of the chapter.

What Richardson saw hidden in the arrows

The most cited warning about causal loop diagrams came from inside the field, from one of its most careful practitioners. Richardson (1986) set it out in a short paper in the field's own journal, with a title that named the problem directly. His argument had two parts. The first was that causal loop diagrams hide the distinction between stocks and flows: a waiting list and a

referral rate appear on the page as the same kind of thing, two words joined by an arrow. The second was that link polarities become ambiguous exactly where accumulations occur. Because accumulation is the idea on which Part I of this book rests, the second point is worth working through slowly.

Return to the hypothetical clinic and add an arrow from "referrals per week" to "waiting list". More referrals mean a longer waiting list than there would otherwise have been, so the link is correctly drawn as positive. Now suppose, again with invented figures, that the clinic receives 120 referrals a week and sees 100 patients a week, so the list grows by 20 a week. A campaign to reduce inappropriate referrals succeeds, and referrals fall to 110 a week. What happens to the waiting list?

The positive sign invites a quick answer: referrals went down, the link is positive, so the waiting list goes down. That answer is wrong. With 110 arriving and 100 leaving, the list still grows, now by 10 a week instead of 20. It will be longer next month than it is today. The positive sign is correct only under the full definition. The list is lower than it would otherwise have been, because without the campaign it would have grown by 20 a week. The reader who drops the counterfactual clause and reads the sign as "moves in the same direction" gets the behaviour backwards.

This is the correlation heuristic that Chapter 2 examined, the assumption that a stock follows the pattern of its inflow, which Cronin, Gonzalez and Sterman (2009) found in well-educated adults across a series of experiments. The uncomfortable point is that the standard notation encodes the heuristic. A positive arrow from a flow into a stock looks exactly like a positive arrow between two ordinary variables, and the ordinary reading of a positive arrow is "they move together". The diagram does not merely fail to correct the error that Part I documented. It presents the structure in a form that invites the error. I think this is the strongest single reason to be cautious about causal loop diagrams as a basis for inference, and it was pointed out by a system dynamicist, not by an outside critic.

The outflow side has the same problem in mirror image. The arrow from "patients seen per week" to "waiting list" is negative. If the clinic has a bad week and sees 90 patients instead of 100, the list is longer than it would

otherwise have been, which is what the minus sign says. But if the clinic improves from 100 to 115 patients a week while referrals stay at 120, the list continues to grow, only more slowly. A manager reading the diagram and the improvement together might expect the list to shrink, and be puzzled, or blamed, when it does not.

Hiding stocks has a second cost beyond the polarity trap. Stocks are what give a system memory and inertia. The waiting list today is the accumulated difference between every referral and every appointment since the list began. That is why, in the clinic example, the list keeps growing for some time after the balancing loop has engaged and extra sessions have been booked: the sessions reduce the net inflow, but until the outflow exceeds the inflow, the stock continues to rise. Stocks also decouple flows, so that the rate at which patients arrive and the rate at which they are seen can differ for long periods, with the list absorbing the difference. A causal loop diagram shows none of this. It can mark a delay on an arrow, but it cannot show that the delay arises because something is piling up, or how much has piled up already.

Two diagram types, and what each gives up

System dynamics has always had a second kind of diagram, and the relationship between the two is the subject of a critical history by Lane (2008). In a stock-flow diagram, stocks are drawn as boxes and flows as pipes with valves that fill or drain them, with information links showing what controls each valve. The waiting list becomes a box; referrals become a pipe flowing in; patients seen become a pipe flowing out. The polarity trap disappears, because nobody looking at a box with two pipes is tempted to think the level moves with the inflow alone. The diagram makes you ask what fills the box and what empties it.

Lane's account weighs what each type gains and loses, both for communicating with the people who have to act on a model and for drawing inferences about how the system will behave. The causal loop diagram is compact. It can be drawn quickly with a group of people who have never heard of system dynamics, it shows feedback at a glance, and it can hold soft variables

such as trust, morale or pressure without anyone having to decide what units they are measured in. Those are genuine advantages, and they explain why the causal loop diagram dominates in workshops. The stock-flow diagram is more demanding. It forces decisions about what accumulates and in what units, it is harder for lay participants to read, and it quickly becomes cluttered. In return it maps directly onto the equations of a simulation model, and it prevents the polarity error.

Neither diagram, on its own, tells you how the system will behave over time. That requires numbers: initial values for the stocks, the strength of each link, the length of each delay. This is the point at which the standard textbook in the field moves from diagramming to formulating and simulating a model (Sterman 2000), and the reason is precisely that people, including trained modellers, cannot reliably work out the behaviour of even a few interacting loops in their heads. Parts of Chapter 3 and Chapter 5 made the same point from the experimental side. Every player of the beer game can see the whole supply chain laid out on the board, and seeing it does not stop the oscillation.

There is a further problem that the two diagram types share. Different people drawing the same system produce different diagrams. Schaffernicht and Groesser (2011) developed a method for comparing mental models of dynamic systems, which is a way of treating diagrams as data about what people believe rather than as pictures of the system itself. The need for such a method is itself informative. If two competent analysts can draw the same clinic with different loops, different polarities and different stocks, then any single diagram is one hypothesis among several, and the choice between them has to be made on evidence rather than on how convincing the picture looks.

The implication for the leverage illusion is direct. A lever is a claim that acting at one point will change behaviour somewhere else in a desired way and by a worthwhile amount. A causal loop diagram contains no amounts and, as Richardson showed, can mislead about direction wherever a stock is involved. It can suggest where to look for leverage. It cannot say whether the leverage is there.

The Foresight obesity map, read as a diagram

The best-known causal loop diagram produced for a government is the obesity system map built for the UK Foresight programme's project on obesity. Its policy story, including the figures in the project's report and what happened afterwards, belongs to Chapter 17. Here it is read only as a diagram, as an example of what the notation can and cannot carry at the largest scale at which it has been used.

The methodology is documented in a separate Foresight paper (Vandenbroeck, Goossens and Clemens 2007). The map was built in workshops with experts, and it was organised around a central engine of energy balance, the relation between the energy people take in and the energy they expend, with the many other causes arranged around that core. It is a qualitative map in causal-loop style, not a simulation model. The project report drew from the wider work the judgment that obesity is shaped by a complex web of causes "where no single influence dominates" (Butland et al. 2007).

The map deserves credit on several counts, and I want to give it before turning to the limits. It put the breadth of causes in front of a government audience in a form that was hard to reduce to a single villain, whether individual willpower, the food industry or the built environment. And it gave systems language a prominent place in a government report, which is part of why it became so widely cited and why it is the reference point for much of what Chapter 17 describes. Finegood, Merth and Rutter (2010), for example, used the map to argue that solutions would need action at several different levels of intervention, an argument that borrowed its levels from the Meadows essay examined in Chapter 12.

Read with Richardson's warnings in mind, though, the map shows the limits of the notation very clearly. The first concerns its central claim. The judgment that no single influence dominates is a statement about the relative sizes of effects. A causal loop map contains no sizes, so it cannot establish that judgment; it can only display a structure consistent with it. The map illustrates the conclusion, and the conclusion itself rests on the experts' reading of the evidence behind the arrows. That is a legitimate basis for a judgment, but it is a different basis from the one the picture seems to

offer. A viewer looking at a dense web naturally infers that no single strand matters most, when the density of the drawing reflects how many relationships the workshops chose to record.

The second limit is the stock check, and here the map's own core supplies the example. Energy balance is a net flow: intake minus expenditure. What accumulates is body mass. Chapter 1 described Hall and colleagues' metabolic model, in which weight responds to a persistent change in energy balance only slowly, taking about a year to move halfway to its new level (Hall et al. 2011). A causal loop diagram drawn around an energy-balance engine shows arrows into and out of that engine, but it does not show that the thing policy is ultimately judged by, the prevalence of obesity in a population, is a stock of accumulated history in each body and, at population level, a stock with its own inflows and outflows. An intervention that successfully reduced the net inflow would produce a slow bend in the curve rather than a fall, and on the map nothing prepares the viewer for that.

The third limit concerns how people look at such maps. In a large diagram, some variables have many arrows going in and out. They look central, and it is tempting to treat them as the places to act. But the number of arrows attached to a variable measures how many relationships the mapmakers recorded, not how strong any of them is. A variable with two strong links can matter more than one with 12 weak ones. Nothing in the notation distinguishes the two cases. This is the leverage illusion in its most literal form: the eye finds a hub, and the hub is read as a lever.

These limits are properties of the notation, and they would apply to any map built this way; they are no reflection on the Foresight team's decision to build one. Whether the map changed policy is an empirical question, and Chapter 17 takes it up with the evidence it deserves.

A loop-drawing protocol with a stock check

What follows is a procedure for drawing causal loop diagrams that keeps their strengths and guards against the problems described above. It rests on the warnings of experienced modellers, principally Richardson (1986) and

Lane (2008), and on the standard modelling sequence in Sterman (2000). It has not itself been tested for whether it improves decisions, so it should be treated as expert guidance with a sound rationale, not as a validated method.

1. **Start with the behaviour over time.** Before drawing any arrow, sketch a graph of the problem behaviour over time: the waiting list over the past two years, for instance, and what you fear or hope it will do. The diagram's job is to offer a hypothesis that could explain that graph.
2. **Name variables as quantities.** Use noun phrases that can take "more of" or "less of", and write units beside each one. Replace actions and moods with the quantities they refer to.
3. **Write the sentence behind each sign.** For every link, write out the full polarity sentence, including "than it would otherwise have been". If the sentence sounds wrong when spoken aloud, the sign or the variable is wrong.
4. **Run the stock check.** Go through every variable and ask whether it accumulates. A useful test is whether you could still measure it at a single instant if every flow stopped. A waiting list passes; a referral rate does not. Mark each stock clearly, perhaps by boxing it, and identify its inflows and outflows by name.
5. **Redraw the arrows into each stock as flows.** Wherever an arrow points into a stock, check that it comes from a flow, and add a note that the stock rises whenever inflow exceeds outflow, whatever the direction of recent change.
6. **Sketch one stock-flow fragment with numbers.** For the most important stock, draw the box and its pipes, put illustrative figures on them and step the arithmetic forward for a few periods. Check whether the story the loops tell survives the arithmetic.
7. **Mark delays and guess their lengths.** A rough estimate, such as "about six weeks to add sessions", is better than an unlabelled mark, because it can be checked.

8. **Name the loops, then state what would make each dominate.** For every loop, write one sentence on the conditions under which it would control behaviour and what data would tell you whether those conditions hold.
9. **Record every apparent lever as a hypothesis.** List the places the diagram suggests acting, and for each one note what evidence, simulation or trial would be needed before recommending it.

Applied to the hypothetical clinic, the protocol changes the conversation in a specific way. Step 1 asks what the waiting list has actually done, which grounds the discussion in data. Step 4 identifies the waiting list as a stock, which immediately reframes the referral campaign: it will slow the growth of the list, not reverse it, unless the clinic sees more patients each week than it receives. Step 6 puts numbers on that, and shows how large the gap between inflow and outflow has to be, and for how long, before the list shrinks to an acceptable size. Step 8 forces the team to say when the reinforcing loop through fatigue and rework might overwhelm the balancing loop through extra sessions, and to name the evidence that would warn them, such as rising rework or rising sickness absence. Step 9 turns the diagram's tempting conclusions into a short list of things to test.

The protocol does not make the diagram any less useful for what it does well. The loops still show feedback at a glance, and the diagram can still be drawn with a room full of people in an afternoon. What changes is the status of the picture. It becomes a record of what the group currently believes about structure, annotated with the checks that belief has passed and the ones it has not, and that is a more honest thing to put on the wall than a picture that appears to show where to push. Chapter 11 takes up what happens when the room drawing the diagram becomes the main point of the exercise.

Chapter 11. Building Models Together: What Group Model Building Delivers

A review of group model building published at the start of this century collected 107 reported cases, and its most striking conclusion concerned the reports rather than the cases. Rouwette, Vennix and van Mullekom (2002) found that the way projects were described and assessed was so uneven that firm conclusions about effectiveness were hard to draw. They offered cautious positive conclusions and asked the field to adopt standard outcome measures. Fourteen years later a second review, of at least 45 studies, found the evidence clearer on some questions and still silent on others (Scott, Cavana and Cameron 2016). Between them, the two reviews describe a method that does something real and has rarely been asked to prove the larger things claimed for it.

This chapter follows the method in the order a practitioner would meet it. It begins with why system dynamicists started building models with groups rather than for them, describes the scripts that turned a craft into something repeatable, sketches the steps of a typical project and follows the method into community settings. It then audits the evaluations. Group model building earns clear credit in this book for what it does inside the room. The audit is about the distance between that and what happens after everyone goes home.

Why build the model in the room

A system dynamics model can be built by an analyst working alone from interviews, documents and data. Group model building is the alternative in which the model, or at least its structure, is built directly with the people who own the problem, in facilitated sessions. Andersen, Vennix, Richardson

and Rouwette (2007) describe it as a bundle of techniques rather than a single procedure, and they place it alongside problem structuring and decision support, which is a useful way of seeing its three different aims.

The first aim is problem structuring. Many of the problems brought to modellers are not well defined when they arrive. People in the same organisation disagree about what the problem is, which variables matter and where the boundary of the system should be drawn. Building a model together forces those disagreements into the open, because a diagram cannot be drawn without deciding what goes on it. In this respect group model building is a relative of the problem structuring methods that Chapter 9 described, and it shares their assumption that the people involved see the situation differently and that the differences are part of the problem.

The second aim is learning. Vennix's methods book is subtitled "Facilitating Team Learning Using System Dynamics" (Vennix 1996), and the emphasis is deliberate. The model is partly a vehicle for changing what participants understand about the feedback structure of their own situation, which, given the evidence in Part I about how poorly people reason about accumulation and delay, is an ambitious goal.

The third aim is decision support: producing a model, sometimes a simulation, that helps the group compare policies before choosing among them. This is the aim closest to the leverage idea. A group that has built a model together is in a position to ask where the model says it should act.

Those three aims are distinct, and much of the confusion about whether group model building "works" comes from running them together. A session can succeed at structuring a problem and building shared understanding while producing a model too rough to support any particular decision. Another can produce a technically sound simulation that the group does not trust because it did not understand how the model was built. Evaluation has to say which aim it is measuring, and as the audit below shows, most evaluations have measured the first two.

Scripts: turning a craft into something repeatable

In its early form, group model building depended heavily on the skill of individual facilitators, which made it hard to teach and harder still to evaluate. Two sessions run by different people under the same name could involve quite different activities. One response, from practitioners in the Albany group, was to write the activities down. Andersen and Richardson (1997) described a set of scripted small-group techniques used in their own model-building work with clients, each one a documented procedure for a particular part of a session.

A script, in this sense, is a written recipe for one piece of facilitated work. It says what the activity is for, what the group does, what the facilitation team does, roughly how long it takes and what it should produce. A session is then assembled from several scripts in sequence. The idea was later extended into a shared, documented collection of scripts intended as a collaborative planning tool, so that sessions could be designed deliberately and repeated by others (Hovmand et al. 2012).

The value of this move is easy to underestimate. Scripts make it possible to plan a session in advance and to explain the plan to a client. They make it possible to train new facilitators on something more definite than apprenticeship. And, most importantly for this chapter, they make it possible in principle to evaluate the method, because an evaluation can say which scripts were used and a later study can use the same ones. A method that cannot be described precisely enough to be repeated cannot accumulate evidence, and the scripting work was a direct answer to that problem. I have not reproduced the published scripts or their descriptions of the facilitation roles here; they are documented in the sources and are better read there in full.

The shape of a project, step by step

What follows is a generic outline of how a group model building project is often structured. It is a composite drawn from the logic of the methods literature described above, not the published protocol of any particular team, and real projects vary a great deal.

The first step is scoping, done with a small number of people from the client organisation before any workshop takes place. Its job is to agree the problem to be addressed, who should be in the room and what the client hopes to do with the result. Getting the membership of the group right matters more than almost any technique, because the model can only contain knowledge that someone present is able to contribute, and its conclusions will only be acted on if the people who control the relevant decisions were part of building it.

The second step is eliciting the problem as behaviour over time. Participants are asked what has been happening, in the form of graphs they sketch of key quantities over months or years, and what they fear or hope will happen next. This anchors the work in a pattern to be explained rather than in a list of complaints, and it echoes the first step of the system dynamics modelling sequence that Chapter 8 described.

The third step is eliciting variables and building causal structure. The group names the quantities that seem to drive the behaviour and then connects them, usually as a causal loop diagram drawn on a wall. This is where the diagramming discipline of Chapter 10 matters most. A room of enthusiastic participants can produce a diagram quickly, and it can reproduce every error Chapter 10 described, particularly the polarity trap around stocks, with the added authority that comes from collective agreement.

The fourth step, when the project goes that far, is formalisation. Between workshops the modelling team converts the group's structure into a stock-flow model with equations, identifies the gaps where numbers are missing, and returns to the group with questions and a working simulation. The fifth step is policy testing, in which the group uses the model to compare options and to see which of its intuitions survive. The final step is planning action, in which the group agrees what it will do, who will do it and how progress will be tracked.

Many projects stop at the third step. The group builds a qualitative diagram, discusses it, and moves directly to action planning without a simulation. That is a legitimate choice when the purpose is structuring and shared understanding, and it is also the point at which the method is most exposed

to the leverage illusion, because a qualitative diagram is being used to choose where to act without the step that could test whether acting there will work.

Taking the method to communities

More recently, group model building has moved well beyond client teams in firms and government agencies. Community-based system dynamics, set out in a methods book by Hovmand (2014), adapts the approach for work with communities, in which the participants are residents, practitioners and local organisations rather than a management team, and in which the model is meant to serve the community's own purposes.

The shift changes what the method is for. A methods paper on community group model building for nutrition inequities in New Zealand focuses on engaging participants and mobilising action (Gerritsen et al. 2020). The emphasis in that framing falls on who takes part and what they go on to do, rather than on the technical quality of the model. Reflections from Amsterdam's LIKE project on childhood overweight and obesity describe the effort to apply participatory system dynamics consistently within a local prevention approach, which is itself a sign of how easily the method's principles can be diluted in practice (Luna Pinzon et al. 2025). In rural Australia, group model building with a community produced a causal loop diagram of the causes of childhood obesity, intended to guide local action (Allender et al. 2015). That diagram became part of a larger intervention whose trial Chapter 17 examines in full.

Participatory modelling has also been used to build simulation models for decisions with very high stakes. Occhipinti et al. (2021) developed a participatory system dynamics model of youth suicidal behaviour for an Australian region, and the model projected that a combination of interventions would have the largest effect over 20 years. The work is careful and the purpose is serious. Its output, however, is a set of projections from a model, not observed outcomes, and the difference matters for anyone tempted to cite the projected effect as evidence that the chosen combination works.

Moving the method into communities also changes its politics in ways that matter for evaluation. In a firm, the client who commissions the work usually has authority over the decisions the model is meant to inform, so the route from workshop to action is short, at least on paper. In a community, authority is spread across agencies, councils, schools, businesses and residents, few of whom answer to one another. A model built with such a group cannot be implemented by order. It can only be taken up, piece by piece, by people who choose to act on it. That is one reason the community literature talks about engagement and mobilisation rather than decisions, and it suggests a sensible set of intermediate outcomes to count: who took part, which actions participants actually started, and whether the network of people working on the problem changed. Those are measurable, and they sit between agreement in the room and change in the population.

These community projects share a pattern that is worth stating plainly. Each paper describes a process, or reflects on applying one, or reports what a model projects. None of them, on its own, evaluates whether the group model building produced better outcomes than would otherwise have occurred. That is not a fault of the papers, which do not claim to be outcome evaluations. It does mean that the growing number of community applications shows how widely the method has been adopted while saying little about its effects.

Auditing the evaluations

The two main reviews of group model building effectiveness are the best place to see what is known. They are worth describing with some care, because their conclusions are often cited more strongly than they were written.

The first, by Rouwette et al. (2002), gathered 107 cases of group model building. Its central difficulty was that the cases had not been reported in comparable ways. Some described outcomes in detail and others hardly at all; some assessed participants' reactions and others the fate of the resulting decisions; very few used the same measures. The reviewers drew cautious positive conclusions from what was available, and their main recommenda-

tion was that the field adopt standardised outcome measures so that future cases could be compared. It is a candid review from within the field, and its caution was well judged.

The second review deserves a closer look at what it actually did. Scott and colleagues examined at least 45 studies of group model building published since 1987. The studies used a mix of designs: controlled experiments, field experiments, pretest-posttest designs and observational studies. The reviewers organised the reported outcomes by type and found them strongest for participants' behaviour and learning, which in their framing covers communication, insight, consensus and commitment. They found some evidence that group model building outperforms traditional facilitation on communication and consensus. They also reported two gaps plainly. Group model building had not been compared with other problem structuring methods, and there was little evidence to guide the choice between different group model building tools. Their recommendations for future research were to move from single cases to multiple cases, from controlled settings to applied ones, and from surveys alone to surveys supplemented by objective measures.

It helps to be clear about what each of those designs can and cannot show, because a review that pools them inherits the limits of each. A controlled experiment compares a group model building condition with some alternative under conditions the researcher sets, so it can attribute a difference to the method, but the more controlled the setting, the less it resembles a real organisation facing a real problem. A pretest-posttest design measures participants before and after a workshop, and it can show that something changed, but without a comparison group it cannot separate the effect of the method from the effect of spending two days thinking hard about a problem with colleagues, which might produce some of the same gains under any facilitation. An observational study reports what happened in a project, often through participants' own accounts, and it can reveal mechanisms and surprises, but it cannot say what would have happened otherwise. The reviewers' recommendation to move from controlled to applied settings and to add objective measures is, in effect, a request to combine the strengths of these designs and reduce their separate weaknesses.

The unevenness that Rouwette and colleagues found also has a practical meaning for anyone reading a case report today. A published account of a group model building project typically describes the setting, the sessions and the model, and then reports that participants valued the process and gained insight. When reading such an account, it is worth asking a short set of questions. Who was measured, and how soon after the workshop? Was anything measured before the workshop, so that change can be seen? Was any outcome recorded that did not come from participants' own judgment? Was there any comparison, even an informal one, with a group that did not take part? And does the report say what happened to the decisions afterwards? Most reports answer the first question and few answer the last, which is the unevenness in miniature.

Put together, the reviews support a clear and limited conclusion. I'd grade the claim that group model building improves shared understanding, communication and consensus among participants as B: the finding is consistent across two reviews spanning several decades and a range of designs, and there is some comparative evidence against ordinary facilitation, but most of the outcomes are self-reported and many come from single cases. The claim that group model building leads to better organisational or policy outcomes months or years later gets a D, because almost nothing in either review links the method to measured downstream results, and the reviewers themselves identify the absence of objective measures and applied comparisons as the main gap.

The split grade is the point, and it should be read in both directions. The B is a real endorsement. Communication, consensus and commitment are not trivial outcomes. Anyone who has watched a policy fail because the people who had to implement it never agreed on what the problem was will recognise that a method which reliably produces agreement is valuable, and the evidence that it does so is better than for most of the methods in this part of the book. The D does not mean that group model building fails downstream; it means that nobody has measured its downstream effects properly. Those are different states of knowledge, and a practitioner who treats absence of evidence as evidence of failure is making the mirror image of the mistake made by one who treats consensus as proof of success.

Two further features of the evidence deserve mention. The first is the missing comparison with other problem structuring methods. Chapter 9 described a family of such methods, all of which aim to help groups structure messy problems, and one of that tradition's own advocates has put those methods "in the dock" over their thin evaluation record (Ackermann 2012). Without head-to-head comparisons, there is no basis for saying that group model building builds consensus better than a well-run soft systems study or a cognitive mapping exercise would. The second is the lack of evidence for choosing between tools within group model building. Even the scripts, the field's main contribution to repeatability, have not been compared with one another for effectiveness.

What agreement is worth

The distinction between the B and the D bears directly on the leverage illusion, because it is in group model building that the illusion becomes social. A room of people who have spent two days building a map together, who now share a vocabulary and a picture and who have agreed on where to act, feels as if it has found the lever. The feeling is real, and it is produced by something real: the group has converged. What the feeling cannot tell anyone is whether the group converged on something true.

There are specific reasons to be careful here. The first comes from Part I. The experimental evidence shows that well-educated adults, including professionals in the relevant domain, misjudge accumulation and delay in consistent ways. A group composed of such people does not become immune by being in the same room. It may simply agree on the same error, and the polarity trap described in Chapter 10 is exactly the kind of error that a group can share without anyone noticing, because everyone reads the arrow the same intuitive way.

The second reason is that the measures used to establish the B are mostly participants' reports of their own learning and agreement. Chapter 5 described evidence that experience with a dynamic task can improve people's results without producing insight, and more generally that practice can make people feel they have understood a structure while their mental

models remain largely unchanged. A participant who says after a workshop that they now understand the system better may be right. They may also be reporting the familiarity that comes from having helped draw the picture. A self-report cannot distinguish the two, which is why Scott and colleagues asked for objective measures.

The third reason is that the facilitation team is not neutral. The modellers choose the scripts, shape the diagram as it emerges and, in projects that reach simulation, decide how the group's structure is turned into equations. A group can converge on the modellers' view of the problem as readily as on its own. Nobody need be manipulating anyone for this to happen. It is a feature of any facilitated process, and it is one more reason why agreement in the room is weak evidence about the accuracy of the model.

A hypothetical case shows how these reasons can combine. Consider a regional health service, invented for this illustration, that runs a group model building project on crowding in its emergency department. Clinicians, managers and ambulance staff build a diagram together over two days. It shows a balancing loop through which crowding prompts extra staffing, and a reinforcing loop through which crowding delays discharges from wards, which blocks beds, which increases crowding further. The group agrees that the lever is faster discharge, and everyone leaves energised. The consensus is genuine and the reinforcing loop may well be real. But nobody in the room has asked what accumulates. Patients waiting for a bed are a stock, fed by admissions and drained by discharges, and a faster discharge process that only matches the current admission rate will stop the queue growing without shrinking it. If the service's admissions are also rising, faster discharge may simply slow the deterioration. A post-workshop questionnaire would record high consensus and strong reported insight. The indicator a year later might show crowding slightly worse than before, and the group, having agreed on the lever, might conclude that the discharge programme had been poorly implemented rather than that the diagram had left something out. Nothing in this composite requires anyone to have behaved badly. It only requires the ordinary errors that Part I documented, shared among well-intentioned people, and measures that capture agreement but not accuracy.

None of these reasons undermines the value of agreement for its own sake. A group that has agreed on a flawed map is often better placed to act, learn and revise than a group that has agreed on nothing, provided it treats the map as a hypothesis. The risk lies in the step from agreement to confidence. When a group that has built a model together tells a funder or a minister that its analysis has identified the lever, it is usually reporting its consensus, and the evidence behind the lever itself is typically the diagram and the group's shared judgment. That can be a sensible basis for trying something. It is not a basis for claiming to know that it will work, and the way to close the gap is to measure what happens.

Two workshops and a measurement plan

The practical response to the audit is for practitioners to generate their own evidence. What follows is a design for a two-workshop group model building project with measures built in from the start. It is a design proposal based on the gaps identified by the reviews, not a published protocol, and the measures it suggests are simple enough to collect without research funding.

Before the first workshop. Ask every participant, individually and in writing, three things. First, sketch a graph of how the main problem quantity has behaved and what you expect it to do over the next year. Second, list the three actions you think would most improve it, and estimate roughly how much each would help. Third, answer one short stock-flow question about the problem, built on the logic of the tasks in Chapter 1, such as what happens to a backlog of open cases when new cases fall this quarter but still outnumber the cases closed. Collect these before anyone has seen anyone else's answers.

Workshop one: structure. Elicit behaviour over time, then variables, then causal structure, using the Chapter 10 protocol, including its stock check. End the session by recording the group's shared diagram and its list of candidate actions, with the group's own estimate of how much each would help.

Between the workshops. The modelling team checks the diagram for stock-flow errors, builds at least a small stock-flow model of the central stock with rough numbers, and identifies which of the group's candidate actions the model suggests would help, which it suggests would not, and which it cannot judge.

Workshop two: testing and commitment. Present the model's results, including where they contradict the group's intuitions. Agree actions, owners and indicators, and record the group's predictions: what each indicator will show in six and twelve months if the chosen actions work.

Immediately after the second workshop. Repeat the three individual measures from before the first workshop. Add a short questionnaire on perceived consensus and commitment, since these are the outcomes the reviews support and it is worth confirming them locally.

At six and twelve months. Check whether the agreed actions were carried out, and compare the indicators with the predictions recorded at the second workshop.

This design produces four kinds of evidence, each answering a different question. The comparison of individual diagrams and action lists before and after shows whether participants converged, and methods exist for comparing mental models of dynamic systems systematically (Schaffernicht and Groesser 2011). The stock-flow question, asked twice, shows whether understanding of accumulation actually changed, as distinct from agreement; the taxonomy of systems-thinking skills proposed by Stave and Hopper (2007) can help decide which level of understanding to target, from recognising interconnections up to testing policies in a simulation. The consensus questionnaire confirms the proximal benefit. And the follow-up at six and twelve months is the beginning of downstream evidence: were the agreed actions implemented, and did the world behave as the group predicted?

None of the analysis requires statistical expertise. For convergence, count how many variables and links appear in most participants' individual diagrams before and after, and note which disagreements disappeared and which remained; a disagreement that survives two workshops is often the

most useful thing the project has found. For understanding, compare the proportion of participants who answer the stock-flow item correctly before and after, and look at the reasoning they give, since a correct answer reached by a correlational argument is luck rather than learning. For consensus, the questionnaire scores speak for themselves, but they are most informative when set beside the convergence data: high reported consensus with little actual convergence in individual diagrams suggests that people agreed to the group's picture without adopting it. For the follow-up, write down the predictions before the second workshop ends, in numbers where possible, and compare them with what the indicators show. A group that predicted a fall in its indicator within six months and saw it continue to rise has learned something about its model that no questionnaire could reveal.

Two additions would strengthen the design considerably. If the organisation is running a similar exercise elsewhere with ordinary facilitation or a different problem structuring method, applying the same measures to both creates the kind of comparison the reviews found missing. And if the project is one of several, recording the measures in the same way across all of them moves the evidence from single cases toward the multiple cases that Scott and colleagues recommended.

The grade for what this design can deliver is the grade of the method itself: B for the proximal outcomes it measures in the room, and, until the follow-up data exist, unknown for the outcomes that matter most afterwards. The difference is that a team using the design will know which of those it has, and will be able to say so. Chapter 12 turns from the room where a map is agreed to the essay that tells readers where on such a map to push.

Chapter 12. Twelve Places to Intervene? Meadows's List and the Search for Levers

A list written on a flip chart

The most quoted framework in systems thinking began, by its author's own account, as an act of exasperation. Donella Meadows describes sitting in a meeting about the new global trade regime, and about how it was likely to make the world work worse. As she listened, she grew convinced that the people around her had no idea how the huge new system they were discussing would behave, and that the remedies being proposed, which she characterised as small adjustments to parameters and weak corrective loops, were hopelessly feeble. She got up, went to the flip chart, turned to a clean page and wrote a list of places to intervene in a system, headed as running in increasing order of effectiveness. The list had nine items. She reports that the room was surprised, herself included, and that she realised at once she had a great deal of explaining to do (Meadows 1999).

The essay that grew out of that moment, "Leverage Points: Places to Intervene in a System", was published by the Sustainability Institute in 1999, and a version of its argument reappeared in her posthumous primer (Meadows 2008). It is short, vivid and generous, and it has become the standard reference whenever anyone in the field talks about where to push. It is also the single text in which the leverage illusion is stated most clearly and questioned most honestly, often in adjacent paragraphs. This chapter reads it closely, puts its central claim on trial and then asks what can be salvaged. The order matters. The essay deserves to be read as its author wrote it before it is judged by what others have made of it.

Its first paragraph is the most famous. "Folks who do systems analysis have a great belief in 'leverage points.'" They are places, she wrote, where "a small shift in one thing can produce big changes in everything." That second phrase is the definition of a leverage point this book uses, and it is worth noticing how much it promises: a small shift, big changes, and everything. In the next paragraph she steps back from the belief and observes that it is also the stuff of legend, the magic password and the secret passage, and then comes the sentence that makes her the first critic of what this book calls the leverage illusion: "We not only want to believe that there are leverage points, we want to know where they are and how to get our hands on them." A writer selling a method does not describe her readers' appetite for it as wishful. Meadows did, in the second paragraph.

She then turns to the lore of her own training. Those who studied with Jay Forrester at MIT, she writes, all absorbed one of his favourite stories. As she retells it, Forrester would say that people know intuitively where leverage points are. Time after time he had analysed a company and located a leverage point, perhaps in inventory policy or in the relationship between the sales force and production. Then he went to the company and found that a great deal of attention was already focused on that point, and that everyone was pushing it "IN THE WRONG DIRECTION!"

The story is doing two jobs at once, and they pull against each other. It affirms that leverage points exist and that analysis can find them. It also says that intuition finds the right place and reverses the sign. If that is true, then no list read intuitively is safe, because the reader will locate the right item and push it the wrong way. The story is therefore an argument for careful analysis of each particular system, and against the idea that levers can be read off a general scheme. Chapter 8 described Forrester's counterintuitive-behaviour thesis and graded it C, because the examples are compelling but there is no base rate of how often intuitive policies actually backfire. The wrong-direction story has the same status. It is a recollection from consulting practice, retold by a student, and nothing in the essay tells us how often it happened or how the right direction was confirmed.

Meadows adds two examples from Forrester's models. The first is the world model, her own introduction to systems analysis, which she says identified growth as the leverage point that world leaders were pushing the wrong way; Chapter 15 examines that model and its critics. The second is Urban Dynamics, which she presents as having shown that less subsidised low-income housing would leave a city better off, a conclusion she describes as derided at the time and later vindicated by the demolition of housing projects. Chapter 8 set out the other side: Kadanoff (1971) argued that the model's counterintuitive conclusions followed from its normative scheme, and that a different scheme reversed them. Meadows's retelling is a loyal student's account, and it presents a contested result as settled.

Then comes a paragraph that is easy to skim past and that contains, I think, the most important caution in the essay. The systems analysts she knows, she writes, have come up with "no quick or easy formulas for finding leverage points." When they study a system, they usually learn where its leverage points are, but a new system is another matter; give them a few months or years, she says, and they will work it out, and even then hardly anyone will believe them. In other words, in her own account leverage is found by sustained study of a particular system, and the finding is hard-won and provisional. What she describes is the modelling of a particular system, which leaves little work for a general list to do. The list she is about to present cannot be the method by which her colleagues found leverage, because by her own testimony they found it by other means.

From nine items to twelve

Meadows is equally candid about the list's origins. She says the flip-chart version was not tightly reasoned, which is characteristic, she adds, of ideas that come to her in a state of boiling over. When she shared it, systems analysts who had their own lists and activists who wanted to use hers immediately sent back questions and comments, and these led her to rethink, add and delete items, change the order and add caveats. Her stated reason for telling this story is to place the list in a context of humility and to

leave room for it to evolve. What follows, she writes, is "a work in progress. It's not a recipe for finding leverage points. Rather it's an invitation to think more broadly about system change."

The two lists are worth setting side by side, in paraphrase, because the differences between them are the best evidence of what she meant by a work in progress. The first list, from the flip chart, ran from least to most effective as follows: constants, parameters and numbers; regulating negative feedback loops; driving positive feedback loops; material flows and the nodes where they intersect; information flows; the rules of the system; the distribution of power over those rules; the goals of the system; and, at the top, the mindset or paradigm from which the system arises.

The revised list has twelve items. From least to most effective: constants, parameters and numbers such as subsidies, taxes and standards; the sizes of buffers and other stabilising stocks relative to their flows; the structure of material stocks and flows, such as transport networks and population age structures; the lengths of delays relative to the rate of system change; the strength of negative feedback loops relative to the impacts they are trying to correct; the gain around driving positive feedback loops; the structure of information flows, meaning who does and does not have access to information; the rules of the system, such as incentives, punishments and constraints; the power to add, change, evolve or self-organise system structure; the goals of the system; the mindset or paradigm out of which the system arises; and, at the very top, the power to transcend paradigms.

Four changes stand out. Two items near the bottom are new: buffers and delays. One item at the top is new: the power to transcend paradigms. And one item was substantially rewritten: the distribution of power over the rules became the power to self-organise system structure, which is a different idea, closer to evolution and innovation than to politics. The fourth change is in the order itself. In the first list, material flows sat above both kinds of feedback loop. In the revision, the structure of material stocks and flows sits below them. Between a first draft and a considered revision, then, a physical category moved from being more effective than feedback to being less effective than feedback, and nothing in the essay reports evidence

that prompted the move. The change came from reflection and conversation. That is a perfectly reasonable way to write an essay. It is not the way a ranking of effects is established, and Meadows never said it was.

Meadows as her own first critic

Before she reaches the list, Meadows pauses to teach the basic grammar of stocks and flows, and she does it with the image Chapter 1 used: a bathtub whose water level rises when the tap runs faster than the drain and falls when the drain runs faster, and which responds sluggishly because flows take time to accumulate. She then complicates it in stages. The bather wants a particular level and adjusts the tap and the drain to reach it, which introduces two correcting loops whose goal is invisible to anyone watching from outside. Then there are two taps, hot and cold, with the boiler four floors down so that the hot water responds slowly, and a bather who is not paying close attention. Then the bathtub becomes a bank account, linked to a thousand others and to a banking system that lends against their combined deposits. Her conclusion is that simple stocks and flows, plumbed together, quickly make systems too complex to work out, and that this is why leverage points are not intuitive. It is a good piece of teaching, and it is consistent with the laboratory evidence in Part I. It is also, once again, an argument against reading levers off a general list, since the complexity she describes belongs to each particular plumbing arrangement.

The body of the essay walks through the twelve places one at a time, and the walk is full of qualifications that the list alone does not carry. A fair reading has to take them seriously, because they show that Meadows understood the problems with her ranking better than many who have used it since.

Take the bottom of the list. Parameters, she argues, attract most of the attention in public argument and rarely change behaviour: they seldom kick-start a stagnant system, stabilise a volatile one or brake a runaway one. But she immediately adds a large caveat. Parameters become leverage points when they move into ranges that trigger one of the items higher on the list; interest rates and birth rates, for example, control the gain around reinforcing loops. She also says that such critical numbers are much less

common than people think, because most systems have evolved or been designed to stay well away from them. That is a careful position. It says the lowest item on the list can become a high-leverage item under particular conditions, and it concedes that knowing whether those conditions hold requires knowledge of the specific system.

The next items up reveal something more fundamental about what the ranking measures. Buffers, she says, can sometimes stabilise a system powerfully, but they are usually physical, expensive and slow to change, so she has not put them high on the list. The physical layout of stocks and flows can have an enormous effect, but rebuilding it is slow and costly, so the leverage lies in designing it properly in the first place. Delays are the clearest case. She says outright that she would list the length of delays as a high leverage point, except that delays are not often easily changeable; things take as long as they take, and it is usually easier to slow the rate of change so that the delays cause less trouble.

Read carefully, these passages show that the ranking combines at least two different questions. One is how much difference a change at this point would make if it could be made. The other is how feasible it is to make the change at all. Buffers and delays are placed low mainly because of the second question. Paradigms, near the top, are placed there on the first question in spite of the second. Meadows sees this. She writes that one could argue paradigms are harder to change than anything else about a system and should therefore be at the bottom of the list, and she replies that nothing about paradigm change is physical, expensive or slow in a single individual, where it can happen in a moment, although whole societies resist challenges to their paradigms harder than anything else. The order of the list, in other words, answers different questions at different heights. This is part of what she meant, I suspect, when she called the order slithery.

The middle of the list is where Meadows's reasoning is most closely tied to a model she knew well. Reducing the gain around a reinforcing loop, she argues, is usually more powerful than strengthening the balancing loops that oppose it, and she illustrates the point with the world model, in which slowing population and economic growth gives the many corrective

processes of technology, markets and adaptation, all of them limited and delayed, time to work. She compares it to slowing a car that is going too fast rather than asking for better brakes. Here, unusually, the ranking between two adjacent items has a basis inside a simulation, and the argument is explicitly about the relation between the speed of growth and the length of delays. Whether the same ordering holds in systems she had not modelled is another matter, and she does not claim it does. The same item introduces the "success to the successful" loop, in which winners gain the resources to win again, a pattern that reappears as one of the archetypes discussed in Chapter 13.

The exceptions she promised in her final caution are visible within the essay too. Replacing the people who run a system is placed near the bottom, alongside parameters, on the reasoning that new hands on the same old taps, following the same information, goals and rules, will not change much. A few pages later she writes that the exception to this is at the top, where a single leader can occasionally change a system's goal and swing large numbers of people in a new direction. The same kind of action, a change of personnel, thus sits at the lowest level or near the highest depending on who is replaced and what they do next. This is a small example of a general difficulty: the list classifies types of change, while the effect of any real change depends on its particulars.

The essay also shows, by its own examples, that the categories overlap. Taxes appear in the first item as examples of parameters. Pollution taxes appear again, later in the essay, among the ways of strengthening negative feedback loops, which is the fifth item from the bottom. The Toxics Release Inventory, discussed below, is presented as a change in information flows, but it was a legal requirement to report emissions, which could equally be classed as a rule. The overlaps are understandable, since a single real intervention usually works through several parts of a system at once, so that where it sits on the list depends on which part of it the analyst chooses to emphasise. A ranking of categories cannot be a ranking of interventions if the same intervention can be placed in more than one category.

The essay ends with a section Meadows titled "A Final Caution", and its sentences are the ones this book has leaned on most. "It is tentative and its order is slithery. There are exceptions to every item that can move it up or down the order of leverage." She adds that years of carrying the list around in her head had not made her superhuman, and then she offers a sentence that, taken seriously, changes how the whole list should be read: "The higher the leverage point, the more the system will resist changing it". She closes by saying that magical leverage points are not easily accessible even when we know where they are and which way to push, and that mastery depends less on pushing leverage points than on "strategically, profoundly, madly letting go."

The resistance sentence deserves more attention than it usually gets. If the potential effect of an intervention rises as one moves up the list, and the resistance to making it also rises, then the list says nothing on its own about where the expected return on effort is highest. A paradigm change might transform a system if it happened, and be very unlikely to happen. A parameter change might do little, and be easy to make. Which is the better use of a given amount of effort depends on magnitudes the list does not supply: how large each effect would be, how likely each change is to succeed and how long it would take. Meadows supplied the principle that forces this conclusion. What she did not do, and did not claim to do, was supply the magnitudes.

It would be unfair to present all this as if Meadows had been naive and her caveats were an afterthought. The caveats run through the essay from its first paragraph to its last. It would also be incomplete to ignore that she headed both lists as running in increasing order of effectiveness, that she wrote confidently in places, and that the numbered format invites exactly the reading she warned against. A numbered list from one to twelve is a ranking in form, whatever its author says about it. The most accurate summary I can give is that Meadows stated the leverage illusion, named the desire behind it, supplied a ranking that feeds it and wrote the most important objections to that ranking into the same essay. She was both the illusion's most quoted source and its first critic, and the second role has been much less remembered than the first.

The meter in the front hall

Two examples in the essay have done more than any others to carry its argument beyond the field, and both appear under the information-flows item. They are the most persuasive passages in the essay, and they are the least supported.

The first is introduced with the words "the story goes". In a subdivision of identical houses, Meadows writes, the electric meters in some houses had been installed in the basement and in others in the front hall, where residents could see them turning faster or slower as they used more or less electricity. With no other change and identical prices, she reports, consumption was 30 per cent lower in the houses where the meter was in the front hall. The essay gives no source for the story, no location, no date, no indication of who measured consumption or how, and no account of how the houses came to have meters in different places.

The second is the Toxics Release Inventory, which Meadows describes as a US government requirement, introduced in 1986, that factories releasing hazardous air pollutants report those emissions publicly every year. There were no fines and no new limits, she writes, only information, and yet by 1990 emissions had dropped 40 per cent. She attributes the continuing decline less to citizen outrage than to corporate shame. Again there is no citation in the essay for the figure.

The absence of sources is striking because the essay is not unreferenced. Meadows gives footnotes for Forrester's books, for her friend who sent her a story about setting a fair rent, for a commodity-cycle model, for Galbraith, for Emerson and for Kuhn. The two numbers that do the most persuasive work in the entire essay are the two that come without one. I do not think this reflects any intention to mislead. The meter story, by its introduction, is offered as lore of the kind that circulates among practitioners, and the essay is written as an essay, not as a research paper. But the consequence is that the book cannot treat either figure as a finding. Both are Meadows's unsourced examples, and they can be reported only as that.

It helps to be specific about what would be needed to turn either story into evidence. For the meter, one would want to know whether the houses really were identical in everything but the meter's position, including who lived in them; whether residents chose the meter's position or had it chosen for them; over what period consumption was measured; and whether the difference persisted. For the inventory, one would want to know whether the fall was in actual emissions or in reported ones, what else changed in regulation and industry over the same years, and what happened to comparable emissions that were not subject to reporting. These are ordinary questions, the kind any evaluator would ask of a claimed effect. The essay does not raise them, and the stories are told in a way that makes them feel unnecessary.

The mechanism the stories illustrate is still worth taking seriously. Meadows's point is that making a flow visible to the people who control it adds a feedback loop that was not there before, and that such additions can be cheap relative to rebuilding physical structure. She also warns that the missing feedback has to be restored to the right place and in a compelling form: telling every user of an aquifer that the water table is dropping might simply start a race to pump first, whereas a price that rises steeply once pumping exceeds recharge would put the signal where it changes behaviour. That caution is subtle and correct as a matter of logic, and it connects to firmer evidence elsewhere in this book. Chapter 2 described experiments showing that the format in which dynamic information is presented changes how well people reason about it, which makes better displays a real, if modest, lever. The meter story is a vivid instance of that idea. Its 30 per cent is the part that has no support.

The list's afterlife, and the missing trial

In the years after the essay appeared, the list acquired a second career in sustainability science, and the way it was taken up is instructive. Very little of the later work tested the ranking. Most of it reinterpreted the list, simplified it or used it as a frame for classifying other things.

The most influential reinterpretation recasts the twelve places as a spectrum running from shallow to deep leverage points. Abson et al. (2017) proposed that reframing and argued that sustainability research had concentrated on tangible but weak leverage points while neglecting the deeper ones. Fischer and Riechers (2019) proposed a leverage-points perspective as a heuristic for sustainability science more generally. Both moves are explicitly conceptual. They keep the ordering while blurring its fine structure into two bands, which makes the scheme easier to use and removes some of the slitheriness Meadows warned about, at the cost of treating the direction of the ranking as settled.

Work carried out for IPBES took a related path. Through a process of expert deliberation, Chan et al. (2020) identified eight leverage points and five levers for pathways to sustainability. The result is a considered expert judgment, drawing on a wide literature, and it should be read as that: the considered view of a panel, not an empirical ranking of interventions by their measured effects. Leventon, Abson and Lang (2021) then offered nine guiding questions for using leverage points in research and practice, and acknowledged that the concept still carries conceptual ambiguities.

A second strand turned the list into a coding frame. Malhi et al. (2009) applied an intervention-level framework derived from Meadows to the background material of a conference on food systems, sorting proposed actions into levels. Johnston, Matteson and Finegood (2014) applied a similar framework to the recommendations in 12 obesity planning documents published between 2004 and 2013, and found that most of the recommendations targeted the lower, less transformative levels. Chapter 10 mentioned Finegood's related argument, built on the Foresight map, that obesity solutions need action at several levels at once.

These are useful studies, and it is important to be precise about what kind of evidence they provide. The coding studies measure where plans and proposals sit on the Meadows scale. They show, convincingly, that obesity plans in that period mostly recommended parameter-level and similar actions. What they do not show is that plans targeting higher levels would have worked better. To describe a predominance of low-level recommenda-

tions as a shortfall already assumes that the ranking is correct. Johnston and colleagues used the hierarchy as a frame, which was their stated purpose, and did not claim to test it. Chapter 17 returns to their study from a different angle, as evidence about what the Foresight map did or did not change in policy.

The claim that needs testing is the one the list is most often used to support: that leverage points can be ranked, so that interventions at higher levels, on goals, rules or paradigms, produce larger effects than interventions at lower levels, on parameters or physical flows. What would it take to test it?

In principle the design is straightforward. Take a domain in which many interventions have been evaluated on a common outcome. Classify each intervention by its Meadows level, using coders who work independently and whose agreement is measured. Compare the effect sizes across levels, allowing for cost, time to effect and the conditions under which each was tried. If the higher levels consistently produce larger effects per unit of effort, the ranking gains support. If they do not, or if the pattern reverses under some conditions, the ranking needs revising.

No such study has been done. The closest thing to a systematic look at the evidence base is the review by Dorninger et al. (2020), which classified empirical studies of interventions in food and energy systems by the depth of the leverage point they addressed. Its main finding was that deep leverage points, which the review took to be those concerning rules, goals and paradigms, are rarely studied at all, and that the research methods in common use shape which interventions get studied in the first place. The empirical literature, in other words, is concentrated at the shallow end of the list, where effects are easiest to measure. That means the evidence needed for a comparison across levels mostly does not exist.

The classification problem alone is enough to make such a study hard, and a hypothetical example shows why. Suppose a city, invented for this illustration, introduces a charge for driving into its centre at peak hours. One coder files it as a parameter, since it sets a price. A second files it as a change to information flows, since the charge puts the cost of congestion in front of each driver at the moment of decision, much as Meadows's meter

put consumption in front of each household. A third files it as a rule, since it alters the terms on which the centre can be used at all. A fourth notes that the revenue is earmarked for public transport, which changes the structure of material flows over the following decade. If the scheme then cuts congestion substantially, which level gets the credit? The answer depends on the coder, and a comparison of effects across levels would inherit that uncertainty. Any serious test would need to show first that independent coders agree on where interventions belong, and I am not aware of any leverage-points study in the sources for this book that reports such agreement.

There are good reasons for the gap, and they are worth stating because they suggest the ranking may be untestable in its general form. Interventions at the higher levels tend to act slowly, diffusely and through many channels at once, which makes their effects hard to attribute and hard to measure against a comparison. Classification is uncertain, as the essay's own examples show, so the same intervention might be coded at different levels by different analysts. And the higher levels are, on Meadows's own account, the ones systems resist most, so an attempt to intervene there may fail for reasons that say nothing about what would have happened had it succeeded.

On this record the verdict is clear. I'd grade the claim that leverage points can be ranked by effectiveness D: the support consists of theory and expert deliberation, the author of the list described its order as tentative and slithery, the evidence base is concentrated at the shallow end, and no study has compared the effects of interventions at different levels. The grade applies to the ranking as a claim about effects. It says nothing against the list as a way of thinking, and it sits alongside two related judgments made elsewhere in the book: Forrester's claim that social systems are counterintuitive, graded C in Chapter 8, and the claim that system maps change policy, graded D in Chapter 17.

It is worth repeating that the D is a judgment on how the list has been used rather than on what Meadows claimed. She offered an invitation to think more broadly. The illusion lies in treating the invitation as a result, so that a recommendation acquires authority from the level it is filed under rather than from any evidence that it works.

Where leverage is real

If the general ranking cannot be defended, the idea of leverage need not be abandoned. There are narrower senses in which small changes do produce large effects, and in which the claim can be checked. Four are worth describing, because together they show what a defensible lever claim looks like.

The first is leverage inside a model whose structure has been tested. In a simulation model, it is possible to vary each parameter and each structural assumption in turn and see which ones the model's behaviour is sensitive to. The standard textbook of system dynamics treats this kind of testing as part of the modelling process (Sterman 2000). The result is a map of where, in that model, small changes make large differences, and it is exactly the kind of knowledge Meadows's parameter caveat calls for, since it can show whether a system is operating near one of the critical ranges she described. The leverage found this way is conditional. It is only as good as the model's structure and data, and it has to be confirmed in the world. But it is a claim with its reasoning attached, and it can be checked, which is more than a position on a list can offer. It is also, as the close reading showed, much closer to how Meadows said her colleagues actually found leverage: by months or years of study of a particular system.

The second is leverage under increasing returns. Chapter 7 described Arthur's formal model of competing technologies, in which, under increasing returns, small historical events can lock in one technology over another, including an inferior one (Arthur 1989). Here the phrase "a small shift in one thing can produce big changes" is literally true. But the conditions are restrictive: the system must have increasing returns, and the small shift has to come early, before one option has pulled ahead. As I read

the model, it is also better at explaining a lock-in after the fact than at identifying in advance which small event will prove decisive, since the events that matter are, by construction, small and historical. The lever exists, but it is time-dependent and hard to see while it can still be pulled.

The third is the counterintuitive model result, which identifies leverage by showing where intuition would push the wrong way. Jones et al. (2006) described a system dynamics model of diabetes population dynamics developed for the US Centers for Disease Control and Prevention (CDC). One of its results was that better control of diabetes, a desirable goal in its own right, can raise the prevalence of the disease, because people with diabetes survive longer. The model also showed that long delays separate primary prevention from its effects on outcomes. This is Chapter 1's stock logic applied to a population. Prevalence is a stock, fed by new cases and drained by deaths, and reducing the outflow raises the stock even as individual health improves. A health system that judged its diabetes programme by prevalence could conclude that success was failure, and a system that wanted to reduce prevalence would need to act on the inflow, with long delays before the result appeared. That is a real insight about levers, found by a model, stated with its mechanism. It is a model result rather than an evaluated outcome, and Chapter 17 discusses the wider question of whether such models have improved health policy.

The fourth is the institutional rule that has been tested across many cases. Chapter 18 describes Ostrom's work on common-pool resources, in which a set of institutional design principles was found in the field and then examined across a large number of cases by later reviewers. On Meadows's list, such principles would sit at the level of rules, fairly high up. What makes them credible is not their level but their evidence: specific rules, under specific conditions, compared across many real institutions. Ostrom found no single lever. She found a set of conditions associated with durable self-governance, and she tested them.

What the four have in common is the thing the general ranking lacks. Each identifies leverage in a particular system or class of systems, states the mechanism, and supplies a way of checking the claim, whether through a

sensitivity test, a formal model with stated assumptions, a simulation with an explicit structure or a comparison across cases. The idea of leverage survives when it comes with its evidence attached. It loses its force when the evidence is replaced by a position on a numbered list.

Prompts first, evidence second

The practical conclusion is that the twelve places are a good way of generating options and a poor way of choosing among them. The two tools below keep those jobs separate. The ranking behind the first tool is graded D as a claim about effects; its usefulness as a prompt does not depend on the ranking being true.

The first tool is a **leverage question set**. Take the problem you are working on and go through the list from bottom to top, writing at least one possible intervention for each place. Some prompts that follow the list's categories:

- Which numbers are people currently arguing about, and is any of them near a range where it would change how a loop behaves?
- Where are the buffers, and are any too small to absorb normal variation or too large to allow a response?
- Which physical arrangements, such as buildings, networks or age structures, lock behaviour in, and could any be designed differently next time they are built?
- Which delays drive overshoot, and can any be shortened, or can the rate of change be slowed instead?
- Which corrective loops are too weak for the pressures they face, and which reinforcing loops are running unchecked?
- Who cannot currently see information that bears on their own decisions, and in what form would it change what they do?
- Which rules, incentives and constraints shape behaviour, and who has the power to change them?
- Could the system be given more capacity to experiment and adapt?
- What goal is the system actually pursuing, as distinct from the one it states?

- What shared assumptions make the current arrangement look natural, and what would someone who did not share them propose?

The output is a long list of candidate interventions across all levels, which is what the list is good at producing. Then discard the ranking. Judge each candidate on four questions: what evidence there is that it would have an effect of useful size in a system like yours; how long the effect would take to appear; how much resistance it would meet, remembering Meadows's warning that resistance rises with leverage; and what side-effects it might have on other parts of the system.

The second tool is a **lever-claim audit**, for use whenever someone, including yourself, claims to have found a lever. Ask:

1. Where did the size of the claimed effect come from: a measured outcome, a model, expert judgment or a story?
2. What was the comparison? What would have happened without the intervention, and how do we know?
3. Over what period was the effect observed or projected, and does it last?
4. What happened to side-effects and to the other stocks in the system?
5. Would another analyst place this intervention at the same level, or does it work through several?
6. Who will resist it, and how?
7. Did the original source attach any caveats, and are they still attached?

Applied to the meter in the front hall, the audit is quick. The effect size comes from a story introduced as lore, with no source. The comparison is between two groups of houses whose other differences are unknown. The duration and side-effects are unknown. The intervention could be classed as information or as physical design. The caveat, "the story goes", was attached by Meadows and is usually the first thing lost when the figure is repeated. The mechanism is plausible and inexpensive to try. The appropriate response is therefore a small pilot with measurement and a comparison group, and the 30 per cent figure should play no part in justifying it.

Used together, the two tools do what Meadows asked of her list and what her own caveats require. The question set widens the search, as her invitation intended. The audit makes sure that whatever is found carries its evidence with it. Chapter 13 turns to the archetypes, where the same idea of reading a lever off a pattern was packaged for managers on a much larger scale.

Chapter 13. The Archetype Library and the Learning Organisation

What would it take to show that an organisation learns, and that learning makes it better at what it does? The question sounds simple, and it is easy to write about organisational learning as if it had already been answered. For many managers, systems thinking arrived through one book rather than through the modelling studies that preceded it, and that book tied systems thinking to a larger promise: that organisations could be redesigned to learn continuously, and that those which did would outperform those which did not.

This chapter takes the book and the movement as two linked cases. The first case is *The Fifth Discipline* and its library of system archetypes, the recurring loop structures that were offered to managers as patterns to recognise in their own organisations. The second is the learning-organisation movement that grew up around and alongside it, with roots that go back further than the book and branches that reach well beyond it. The evidence then arrives, as it would in a courtroom, with witnesses for and against, followed by two witnesses from outside the family whose testimony turns out to matter a great deal. The verdict is mixed, and one part of it is a clear and deserved success.

The book: loops packaged as patterns

Senge's *The Fifth Discipline* (Senge 1990) popularised systems thinking in management by presenting it as one of five disciplines that together would produce a learning organisation, and by giving managers a set of system archetypes to work with. The book drew on the beer game, which Chapter 3 described at length, and on anecdote; it did not present evaluation data, and

it was not trying to. It was followed by a fieldbook of tools and exercises for putting the disciplines into practice (Senge et al. 1994), and by a stream of related writing. Senge and Sterman (1992) set out a programme in which systems thinking, attention to mental models and management flight simulators together formed the route to organisational learning. Kofman and Senge (1993) wrote a normative essay describing communities of commitment as the heart of the learning organisation. An edited volume collected practitioner cases of models used as tools for organisational learning (Morecroft and Sterman 1994), and much later the archetypes were carried into work with foundations and non-profit organisations through a practitioner guide built on causal loops and case vignettes (Stroh 2015).

The archetypes were the part of this body of work that travelled furthest. An archetype is a small causal loop diagram, usually two or three loops, given a memorable name and presented as a structure that recurs across many different settings. Three examples show how they work.

In the archetype usually called "fixes that fail", a problem symptom prompts a fix, and the fix relieves the symptom: a balancing loop. But the fix also has an unintended consequence which, after a delay, makes the underlying problem worse, so that the symptom returns and prompts more of the same fix. The second loop is reinforcing, and because of its delay the fix looks successful for a while before the pattern becomes visible.

In "shifting the burden", a problem can be addressed in two ways: by a symptomatic solution that works quickly, and by a fundamental solution that works slowly. Each is a balancing loop. The symptomatic solution tends to win because its effect arrives first, and it often has a side-effect that erodes the capacity to apply the fundamental solution, so that over time the organisation becomes dependent on the quick fix.

In "success to the successful", two parties compete for a shared pool of resources, and whichever is ahead receives more of them, which helps it stay ahead. Two reinforcing loops are coupled through the allocation, so that a small early advantage can become a large and lasting one regardless of which party had more potential. Meadows used the same name for such loops in the essay discussed in Chapter 12.

Wolstenholme (2003) later argued that the library of archetypes could be reduced to four generic structures built from combinations of problem loops and solution loops. His analysis is a useful corrective to the impression that there is an open-ended catalogue of patterns, each needing to be learned separately. It also makes clear what the archetypes are: compact theories about how an intervention intended to solve a problem can, through feedback, fail to solve it or make it worse.

The appeal is easy to understand, and it is the appeal of the leverage illusion in its most portable form. An archetype arrives with its lever attached. If you recognise your situation as a fix that fails, the diagram itself points to the delayed side-effect as the thing to address. If you recognise shifting the burden, it points to the fundamental solution that has been neglected. Recognition appears to deliver both a diagnosis and a prescription, with no model, no data and no simulation required. For a manager who has absorbed the Part I evidence that people reason poorly about feedback, the archetypes seem to offer a shortcut: learn the patterns, and you will see what others miss.

There is also a reason to expect archetypes to feel more diagnostic than they are. Each one is a story as well as a diagram: a quick fix that backfires, a dependency that deepens, a rich-get-richer spiral. Most organisations contain some version of all three stories somewhere, because most organisations have used quick fixes, relied on short-term expedients and allocated resources by past results. A manager who has just learned the archetypes will therefore find them everywhere, and each recognition will feel like confirmation. The feeling is produced by the fit between a general story and a particular memory, and it is the same whether or not the archetype's loops are actually strong enough to drive the organisation's behaviour. Nothing in the act of recognition distinguishes a true fit from a merely plausible one.

Senge and Sterman's programme had anticipated part of this problem. Their route to organisational learning combined systems thinking with management flight simulators, precisely because a simulation lets people test their mental models against the consequences of their decisions instead of

trusting their reading of a diagram. Chapter 5 described the evidence on such learning environments: structured, interactive experience with a simulator shows gains in several laboratory studies, while unstructured experience does little. The archetypes, as they spread through management teaching, mostly travelled without the simulators. The diagram went everywhere; the test went almost nowhere.

The difficulty is that recognising a pattern and confirming it are different acts. Archetypes are causal loop diagrams, so they carry every limitation that Chapter 10 described. They leave stocks implicit, their polarities can mislead around accumulations, and they contain no magnitudes, so they cannot say whether the harmful loop is strong enough to matter. A manager who recognises a fix that fails has formed a hypothesis about the structure of a situation. Whether the delayed side-effect exists, how large it is and how long its delay lasts are empirical questions that the archetype raises and cannot answer. That becomes the basis of the practitioner tool at the end of this chapter.

The movement: from double-loop learning to diagnostic surveys

The learning-organisation idea is older than Senge's book, and its roots are in a different tradition. Argyris and Schön (1978) developed a theory of organisational learning built around two distinctions. The first is between single-loop learning, in which an organisation detects and corrects errors within its existing goals and assumptions, and double-loop learning, in which it questions the goals and assumptions themselves. The second is between espoused theory, what people say guides their actions, and theory-in-use, what can be inferred from how they actually behave. Argyris (1977) gave the first distinction its most accessible statement for a management audience. The loop language is a metaphor drawn from cybernetics rather than a formal feedback model, but it gave the later movement a vocabulary that fitted naturally with systems thinking.

Kim (1993) proposed a conceptual model linking individual learning to organisational learning through mental models, which connected the Argyris and Schön tradition to the system dynamics interest in how people

represent the structure of their situation. Garvin (1993), writing in the same year, argued that the learning organisation needed clearer meaning, management and measurement if it was to be more than an inspiring phrase, and described a set of building-block activities through which organisations might learn. It is worth noting that one of the movement's own advocates was asking for measurement at this early stage. The request would take years to be met, and when it was, the measurement came in the form of surveys.

The main instrument that emerged was the Dimensions of the Learning Organization Questionnaire, developed by Marsick and Watkins as a diagnostic of an organisation's learning culture (Marsick and Watkins 2003). Yang, Watkins and Marsick (2004) reported a psychometric validation of its seven dimensions, including evidence on model fit, reliability and links to perceived performance. A different diagnostic survey was later proposed by Garvin, Edmondson and Gino (2008), organised around three building blocks, one of which, a supportive learning environment, explicitly includes psychological safety. That link between the learning-organisation tradition and research on psychological safety becomes important later in this chapter.

The double-loop idea also anticipates, in a different vocabulary, the upper reaches of Meadows's list in Chapter 12. Both treat questioning a system's goals and assumptions as a deeper kind of change than adjusting behaviour within them, and both have been widely taken to imply that the deeper kind is more valuable. Neither claim was presented with comparative evidence, and the resemblance is a reminder that the idea of a hierarchy of change, with the most fundamental level at the top, runs through several strands of systems thinking at once. It is an attractive idea. It is also, in each of its forms, a claim about the relative size of effects that has not been measured.

A related strand of thinking, influential on the systems thinking later applied in public services, came from Deming (1982), who treated management as a system, emphasised the understanding of variation and used his red-bead experiment to show how much of apparent individual performance is produced by the system in which people work. Deming's

influence ran in parallel with the learning-organisation movement rather than through it, but it shaped a similar belief: that seeing an organisation as a system reveals where improvement really lies. Not everyone within the systems community accepted Senge's framing. Flood (1999) offered an early critical reappraisal of *The Fifth Discipline* from a critical-systems and complexity perspective.

The case for the learning organisation

The prosecution in a courtroom usually speaks first, but here it is fairer to begin with the defence, because the evidence for the learning organisation is more substantial than its critics sometimes allow.

The strongest witness is a meta-analysis. Goh, Elliott and Quon (2012) pooled 33 studies of the relationship between learning capability, measured in various ways, and organisational performance. They found a positive relationship, and they found that it was stronger for non-financial outcomes than for financial ones. The authors were careful about what their result meant. They flagged the question of causality as unresolved, because the studies they pooled could not show whether learning capability produced better performance or better-performing organisations found it easier to invest in learning.

Individual studies fill in the picture. Ellinger et al. (2002) surveyed managers and linked their organisations' scores on the learning-organisation questionnaire to both perceptual and objective measures of financial performance, and reported a positive association. Kim, Watkins and Lu (2017) used secondary data and a structural equation model to show that learning-organisation culture was related to knowledge performance, which in turn fully mediated its relationship with financial performance. And the measurement work of Yang and colleagues showed that the questionnaire's dimensions hold together statistically, which matters because a correlation between performance and a scale that did not measure anything coherent would mean little.

Taken together, the defence has established something real. Organisations whose members report a stronger learning culture tend to perform better, particularly on non-financial measures, and this association has been found repeatedly with a reasonably well-validated instrument.

The case against the construct

The prosecution's case has two parts. The first is about the design of the supporting studies. The second is about whether the thing being measured is a single thing at all.

On design, every study just described is cross-sectional, and most rely heavily on self-report. That creates three problems that no amount of repetition can fix. The direction of causation is unknown: a successful organisation may have the slack, the confidence and the stability to invest in learning, so performance could produce learning culture as easily as the reverse. The same respondents often rate both the learning culture and aspects of performance, which tends to inflate the relationship between them, because people who feel positive about their organisation are likely to rate it well on both. And a third factor, such as good leadership or a favourable market, could produce both a learning culture and good results without either causing the other. Goh and colleagues acknowledged the causal question directly. No longitudinal or quasi-experimental study showing that becoming a learning organisation raises performance was found in the research for this book.

On the construct, Örtenblad (2002) showed that the phrase "learning organisation" is used in at least four distinct senses in the literature: to mean organisational learning, learning at work, a learning climate or a learning structure. Studies using different senses measure different things, which makes it hard for evidence to accumulate even when individual studies are sound. Caldwell (2012) made a deeper objection from practice theory, arguing that Senge's learning organisation never specified the social practices that would bring it about, and that its systems model has no satisfactory account of human agency. On this view, the learning

organisation describes a desirable end state without saying how people would actually get there. Grieves (2008) went furthest, arguing that the idea had run its course and should be abandoned.

One detail of the defence's own evidence deserves a question from the bench. The association in the meta-analysis was stronger for non-financial outcomes than for financial ones. There are several possible explanations. Learning may genuinely affect things like innovation and satisfaction before it reaches profit, which would be the defence's reading. But non-financial outcomes are also more often measured by asking people, sometimes the same people who rate the learning culture, while financial outcomes more often come from accounts. If that were so in the pooled studies, the pattern would be what common-method inflation predicts: the relationship looks strongest where both variables come from the same respondents. The sources available for this book do not allow that possibility to be checked, and I raise it as a question the evidence leaves open rather than as a finding.

In fairness to the defence, the prosecution's witnesses are not unanimous. Pedler and Burgoyne (2017), responding to claims that the learning organisation was dead, offered a short history and survey findings and argued that the idea survives in altered forms. Their argument is about the idea's continuing use rather than its effectiveness, but it is a reminder that a concept can remain useful for organising practice even when its causal claims are unproven.

Two witnesses from outside the family

Two further witnesses sit slightly apart from the main dispute, and their testimony points in opposite directions.

The first is research on team psychological safety. Edmondson (1999) introduced the construct as "a shared belief held by members of a team that the team is safe for interpersonal risk taking", the kind of risk involved in admitting a mistake or raising a concern. In a multimethod field study of 51 work teams in a manufacturing company, she found that team psychological safety was associated with the team's learning behaviour, and that learning behaviour mediated the relationship between psychological safety and team

performance. The finding has since been examined across a very large number of samples. Frazier et al. (2017) conducted a meta-analysis of 136 samples covering more than 22,000 individuals, and found that psychological safety was related to learning behaviour, engagement and performance.

This deserves full credit, and I'd grade the claim that team psychological safety predicts learning behaviour and performance B: the meta-analytic base is large and the relationships are consistent, but most of the underlying studies are correlational, so causal evidence remains thin, and the size of the effects varies with the measure used. That is a better grade than the learning organisation as a whole receives, and it is worth asking why. The construct is narrower: it concerns a specific shared belief within a team, not the character of an entire organisation. It is defined at a level where it can be measured with some precision. It comes with a proposed mechanism, learning behaviour, that can itself be measured and tested as a link in the chain. And the research programme has been cumulative, with studies using comparable measures so that a meta-analysis could pool them. Psychological safety is, in effect, what the learning-organisation idea looks like when it is made small enough to test, and the Garvin, Edmondson and Gino survey already recognised that by placing it within one of its building blocks.

The finding also matters for the other methods in this part of the book, because several of them quietly depend on it. A group model building workshop, as Chapter 11 described, is valuable partly because it forces disagreements into the open, and that only happens if participants are willing to say that a colleague's arrow points the wrong way or that the official account of a problem is incomplete. Testing an archetype against data requires someone to report that the promotion programme, or the reliance on agency staff, is producing the delayed harm the archetype predicts, which may mean telling senior people that their preferred fix is failing. In a team where speaking up feels risky, the diagram on the wall will reflect what people think it is safe to say rather than what they believe, and the tools will reproduce the organisation's existing account of itself with an added air of analysis. Psychological safety does not make a systems method work, but its absence can quietly prevent one from working.

The second witness has the opposite testimony. Cynefin, introduced by Kurtz and Snowden (2003), is a sense-making framework that challenges assumptions of order, rational choice and intent in strategy, and sorts situations into different kinds of context. A popular version for managers distinguished simple, complicated, complex and chaotic contexts, plus a state of disorder, and proposed a different leadership response for each (Snowden and Boone 2007). Cynefin is a sibling of the archetypes in an important respect. Like them, it offers a way of recognising what kind of situation one is in and implies what to do about it. And like them, it has not been tested. No empirical study was found in the research for this book examining whether using Cynefin improves decisions or outcomes. Its evidential standing is the same as that of the archetypes: plausible, widely taught and untested.

The verdict

On the learning organisation, I'd grade the claim that learning organisations outperform others C: the association with performance is consistent, has been found in a meta-analysis and holds up with a validated instrument, but the evidence is cross-sectional and largely self-reported, the causal direction is unknown and the construct itself is used in several incompatible senses. The fairest summary is that learning culture is a plausible correlate of performance and has not been shown to be a cause.

On the archetypes, the verdict is simpler and harsher. The claim that archetypes diagnose real organisational problems is widely taught and rarely tested, and no study was found that compares archetype-based diagnosis with any alternative, or checks whether managers who identify an archetype in their organisation are right. The support consists of the books that present them, Wolstenholme's theoretical analysis and practitioner vignettes; the principal objections come from within system dynamics itself, in the critiques of causal loop inference by Richardson (1986) and Lane (2008) that Chapter 10 described. This is grade-D evidence, because there are no outcome studies at all.

It is worth being specific about what evidence would change these verdicts, because in both cases the missing studies are feasible. For the learning organisation, the need is for designs that follow organisations over time. A longitudinal study that measured learning culture and performance repeatedly would show which tends to change first. Better still would be an evaluation of a deliberate programme to strengthen learning culture, with a comparison group of similar organisations or units that did not receive it, and with performance measured independently of the people who rate the culture. For the archetypes, the need is for tests of diagnosis. One simple design would present managers with cases whose underlying structure is known, perhaps because the cases come from well-documented simulations, and measure how often they identify the correct archetype, how often they see one that is not there and how often two managers agree. Another would follow managers who have diagnosed an archetype in their own organisation and check, with data, whether the predicted pattern appears. None of these studies would be expensive by the standards of management research. That they appear not to have been done, on the evidence gathered for this book, marks a gap rather than an impossibility.

Psychological safety emerges from the courtroom with the best record, and the contrast is instructive. The part of this family of ideas that holds up best is the one that was narrowed, measured and tested cumulatively. The parts that promised most, a whole organisation that learns and a library of patterns that reveal where to act, are the parts with least evidence behind them.

Archetypes rewritten as hypotheses

The archetypes need not be discarded. They are compact, memorable theories about how interventions go wrong, and that is useful. What they need is to be treated as hypotheses rather than diagnoses, which means stating in advance what evidence would confirm or refute them. The approach below does that for the three archetypes described earlier, using hypothetical examples. It is untested as a method, and it inherits the D grade of archetypes as diagnostic tools; its purpose is to make that grade improvable in a particular case.

Fixes that fail. Consider a hypothetical retailer that uses price promotions to meet quarterly sales targets. The archetype predicts that promotions raise sales in the short run and that, after a delay, customers learn to wait for promotions, which lowers sales at full price and increases pressure for further promotions. If the hypothesis is right, the data should show three things: sales spikes during promotions followed by troughs deeper than the pre-promotion baseline; a rising share of sales made at a discount over successive quarters; and increasing frequency or depth of promotions over time. If full-price sales recover fully between promotions and the discount share is stable, the archetype does not fit, however familiar the story feels.

A little arithmetic, with invented figures, shows how the test separates the two possibilities. Suppose the retailer sells about 1,000 units a week at full price and runs a two-week promotion each quarter. If the archetype is at work, the first promotion might lift weekly sales to 1,600, followed by two weeks at 800 as customers who bought early stay away, before sales settle back at 1,000. Over a 13-week quarter that comes to 13,800 units against 13,000 with no promotion at all, so the first promotions appear to add volume. After a year of such cycles, if customers have learned to wait, the ordinary weeks might settle at 900 instead of 1,000, and the quarter's total would fall to 12,900 units, below the level with no promotions, even though each promotion still looks as strong as the first. If the archetype is not at work, the post-promotion dip is short and shallow, full-price sales between promotions stay at 1,000 year after year, and the discount share holds steady. The two patterns are easy to tell apart once the data are plotted over several quarters, and very hard to tell apart from inside a single quarter, which is where the decision to run another promotion is usually made.

Shifting the burden. Consider a hypothetical hospital that fills nursing gaps with agency staff while its own recruitment and training capacity declines. The archetype predicts that reliance on agency staff grows over time, that investment in the fundamental solution falls, and that the fall in capacity follows, with a delay, periods of heavy reliance on the quick fix. The data to collect are agency hours, training places and permanent hires over

several years. If agency use rises while training and hiring hold steady or rise, the burden is not being shifted, and the problem lies elsewhere, perhaps in demand.

Success to the successful. Consider a hypothetical firm that allocates its development budget between two product lines in proportion to each line's sales in the previous year. The archetype predicts that an early difference in sales, even a small or accidental one, will widen year after year, regardless of the lines' underlying potential. The test is whether allocation actually follows past performance and whether the gap widens faster than differences in market conditions would explain. If allocation is decided on other grounds, or the gap is stable, the archetype does not apply.

For each archetype, then, the steps are the same in outline but different in content. Write down what the archetype predicts in terms of measurable quantities over time. Name the delay it assumes and estimate its length. Identify the stocks involved, since each of these archetypes contains at least one accumulation that the loop diagram leaves implicit: customers' expectations of discounts, the hospital's stock of trained staff, the firm's accumulated product strength. Collect the data for long enough to cover the delay. Decide in advance what pattern would count against the hypothesis, and accept the verdict of the data.

Alongside this, a **psychological-safety check** is worth running wherever a team is asked to learn from its own performance, because the evidence suggests that learning behaviour depends on it. Published measures of team psychological safety exist, and for any serious assessment they should be preferred to improvised questions. For a quick first look, the following questions, which are my own and not a validated instrument, point in the right direction. When someone in this team makes a mistake, is it discussed openly or kept quiet? Do people ask for help without worrying about how it will look? When a junior member raises a concern about a plan, what happens next? Have errors or near-misses been reported in the last three months, and if none have, is that because none occurred? Behavioural signs are often more revealing than answers to questions: the number of issues raised in meetings, the proportion of incidents reported by the people

involved, and whether bad news reaches senior staff early or late. A team in which none of these signs is present is unlikely to learn from an archetype or anything else, which is the most practical thing the courtroom established. Chapter 14 asks whether the skill that all these tools assume, the ability to think in systems, can be taught and measured at all.

Chapter 14. Can Systems Thinking Be Taught, or Even Measured?

Suppose you were asked to certify that the graduates of a course could think in systems. What would you ask them to do? You might give them a bathtub problem of the kind in Chapter 1 and check whether they can work out when a stock peaks. You might ask them to draw a causal loop diagram of a familiar problem and mark its loops. You might hand them a questionnaire asking how often they consider the wider consequences of their decisions. Or you might watch them manage a simulated business and see whether they avoid the oscillations of Chapter 3. Each of these is a reasonable answer. They measure different things, and nobody has shown that a person who does well on one will do well on the others.

That uncertainty sits underneath the whole of this part of the book. Every method examined so far assumes a skill. Causal loop diagrams assume that people can read feedback structure, group model building assumes that participants learn to see their system more accurately, and the leverage question set of Chapter 12 assumes that users can tell a loop from a list. The leverage illusion depends on the same assumption in its strongest form: that seeing structure is a capacity people can acquire, and that once acquired it will guide real decisions. Part I showed that most people lack the capacity by default. Chapter 5 described evidence that some forms of training help, at least in the short term. This chapter asks the questions an educator would ask before building a curriculum around the skill, and it gives each a mark.

What exactly is the skill?

The first question is what is being taught, and the literature offers several answers that overlap without coinciding.

One influential answer came from Richmond (1993), who argued that systems thinking should be taught as a set of core critical-thinking skills and supplied much of the vocabulary still used for them, including dynamic thinking, closed-loop thinking and operational thinking. On this view the construct is a bundle of distinct abilities, each of which could in principle be taught and tested separately. Stave and Hopper (2007) took a similar approach and organised it into a proposed taxonomy of seven levels: recognising interconnections, identifying feedback, understanding dynamic behaviour, understanding stocks and flows, using conceptual models, creating simulation models, and testing policies. They mapped the levels onto Bloom's taxonomy of educational objectives and suggested assessments for each, which made the taxonomy a practical tool for course designers.

Other definitions pull in different directions. Arnold and Wade (2015) reviewed the competing definitions and proposed one built from the skills a systems thinker should demonstrate, with the explicit aim of making systems-thinking education assessable. Monat and Gannon (2015), reviewing a selection of the literature, defined systems thinking as a perspective, a language and a set of tools, which treats it partly as a way of seeing the world and not only as a set of abilities. Cabrera, Colosi and Lobdell (2008) proposed that systems thinking reduces to four patterns of thought: making distinctions, identifying systems, recognising relationships and taking perspectives. Their framework was aimed at evaluators, and its empirical support has come mainly from the authors' own programme; a popular book later presented it to a general audience (Cabrera and Cabrera 2015). Doyle and Ford (1998) worked on a narrower but related problem, proposing an operational definition of mental models for system dynamics research, since a skill defined in terms of mental models can only be measured if the models themselves are defined precisely.

The differences matter because they determine what a test can look for. A definition built from performance abilities, such as reasoning correctly about accumulation, leads naturally to tasks with right and wrong answers. A definition built around a perspective or disposition leads naturally to questionnaires about how people habitually approach problems. A framework of four thinking patterns leads to exercises in which people are

asked to apply the patterns. It is quite possible for someone to score well on one of these and poorly on another, and if so, the three communities would disagree about whether that person can think in systems.

Hopper and Stave (2008), reviewing efforts to assess systems-thinking interventions in the classroom, reached the conclusion that follows from all this: there was no agreed definition of systems thinking and no agreed best way of assessing it. Their review is now some years old, but nothing in the more recent literature gathered for this book suggests that consensus has since been reached.

Mark for the first question: incomplete. Several careful definitions exist and some, such as the Stave and Hopper taxonomy, are usable in practice. The field has not agreed which one describes the skill it promises to teach.

How would anyone know a person has it?

Measurement instruments have multiplied, and they fall into two broad families that rarely meet.

The first family is self-report scales, in which people rate themselves on statements about how they think. Davis and Stroink (2016) gave such a measure to 115 undergraduates and found that systems-thinking scores were associated with an ecological worldview, connectedness to nature and pro-environmental behaviour, in a cross-sectional design. Randle and Stroink (2018), across two studies, developed and began to validate a further self-report measure and found that it overlapped with intelligence and cognitive complexity without being fully explained by them. Thibodeau, Frantz and Stroink (2016) located a self-report systems-thinking measure among related psychological constructs, testing which it should and should not resemble. In health care, Dolansky et al. (2020) developed and validated a 20-item Systems Thinking Scale for quality-improvement education. In engineering education, Camelia and Ferris (2018) validated a 16-item questionnaire with a three-factor structure, in samples of 186 and 163 students, designed to measure students' affective engagement with systems thinking.

These are serious pieces of psychometric work, and they show that the scales hold together statistically and relate to other variables in sensible ways. What they measure, though, is how people describe their own habits of thought, or in the last case how they feel about systems thinking. The Davis and Stroink association with environmental attitudes illustrates the difficulty of interpretation: a person who endorses statements about interconnectedness may be expressing a worldview as much as a reasoning skill, and a cross-sectional association cannot separate the two.

The second family is performance tasks, in which people have to do something that can be scored. The systems thinking inventory of bathtub and department-store problems, whose results Chapter 1 described, is the best-known example (Sweeney and Sterman 2000). Plate (2010) developed a cognitive-mapping assessment and found that students who had received systems-oriented instruction understood a novel environmental system better than those who had not. Grohs et al. (2018) built a scenario-based assessment, scored with a rubric, in which students reason about an ill-structured problem, and developed it on 93 student responses. Mehren et al. (2018) used item-response methods to validate a two-dimensional model of system competence in geography education, distinguishing understanding of a system's organisation and behaviour from a system-adequate intention to act.

A systematic review of systems-thinking assessments in engineering found many formats, few that gave equal weight to technical and contextual reasoning, and a need for more performance-based assessment (Dugan et al. 2022). The review's picture is of a fragmented field in which instruments are developed for particular courses and populations and rarely compared with one another.

Two further problems cut across both families. The first is the overlap with general intelligence. Chapter 4 described the evidence on complex problem solving: the overlap with intelligence is substantial, and whether complex problem solving adds anything beyond it depends on how broadly intelligence is measured. The self-report work of Randle and Stroink found a similar partial overlap. Any claim that a systems-thinking measure captures

a distinct skill therefore needs to show that it predicts something beyond what a good intelligence test would predict, and few studies have attempted this. The second problem is convergence. The two families have rarely been administered to the same people and compared, so there is little evidence that a person who rates themselves highly on a self-report scale will also perform well on a bathtub task or a mapping assessment.

The study that would settle the convergence question is not complicated to describe. Give the same group of people two or three different self-report scales and two or three different performance tasks, together with a measure of general reasoning. If systems thinking is a single coherent skill, the performance tasks should agree with one another more than with the reasoning test, and the self-report scales should agree with the performance tasks at least moderately. If the self-report scales agree only with each other, and the performance tasks track the reasoning test, the field would learn that it has been measuring two things under one name: a disposition to value interconnectedness and a general capacity for abstract reasoning. Either result would be useful. At present, the literature gathered for this book does not allow anyone to say which is closer to the truth.

I'd grade the claim that systems thinking can be measured as a coherent individual skill C: many instruments exist and several have respectable psychometric evidence behind them, but there is little evidence that they converge on the same construct, few are performance-based, and the overlap with general ability is unresolved.

Mark for the second question: promising but fragmented. The individual instruments are better than the field's reputation suggests; the relationship among them is largely unknown.

Can children learn it?

The school-level research gives the most detailed picture of what happens when people are taught to think in systems, and it begins with a sobering baseline. Sweeney and Sterman (2007) gave instruments on natural and social systems to US middle-school students and their teachers. Both groups showed limited intuitive systems thinking: explanations tended to run in

open chains rather than loops, delays were missed, and stock-flow misconceptions appeared at all ages. Teachers did better than students, but a quarter of the students scored at the level of the teachers' median. The finding matters for curriculum design because it means teachers cannot be assumed to hold the understanding they are being asked to teach.

Research on how learners understand complex systems more generally points to the same kinds of difficulty. Hmelo-Silver and Pfeffer (2004) compared how experts and novices described an aquarium and found that novices focused on its structures, while experts focused on its behaviours and functions. Jacobson (2001), in a small comparison, found that novices favoured explanations involving centralised control and single causes, while experts accepted decentralised, probabilistic and emergent explanations. Chi et al. (2012) developed an account, supported by studies, of why students misunderstand emergent processes such as diffusion and natural selection: they treat them as direct processes driven by particular agents. Wilensky and Resnick (1999) argued that confusion between the level of individuals and the level of emergent patterns underlies many such misconceptions, and presented agent-based modelling as an educational remedy.

These findings are worth connecting to the argument of the book. The novice preference for centralised control and single causes that Jacobson described is, in miniature, the cognitive signature of the leverage illusion: the expectation that a system's behaviour has one source and can be moved from one place. If that expectation is the default for learners, then teaching systems thinking is partly a matter of dislodging it, and a curriculum that presents systems thinking mainly as a way of finding the key intervention may reinforce the very intuition it should be correcting. The research on emergent processes points the same way. Understanding that a pattern can arise from many interacting parts, with no single part in charge, is precisely what makes the idea of a single lever less plausible, and it is also, on the evidence of Chi and colleagues, one of the hardest things for students to grasp.

Against this background, several intervention studies report gains. Assaraf and Orion (2005) followed about 50 Israeli eighth-graders through an earth-systems curriculum and described systems-thinking skills developing in hierarchical stages. Most students progressed, and about a third reached the highest level. The study had no control group, so it shows that students changed during the curriculum but not that the curriculum caused the change. Gilissen, Knippels and van Joolingen (2020) used a lesson-study design in two upper-secondary biology classes of 26 and 19 students and found that students could name and apply seven characteristics of systems; the study was designed to produce guidelines for teaching rather than estimates of effect. Hmelo-Silver et al. (2017) reported a quasi-experimental classroom study of a middle-school curriculum organised around a conceptual representation of a system's components, mechanisms and phenomena. A review of how systems thinking has been used in science, technology, engineering and mathematics education maps the growing range of such work (York et al. 2019).

The anchor study for this chapter is worth describing in more detail, because its design allows a stronger conclusion than most. Riess and Mischo (2010) ran a quasi-experiment with 424 German sixth-graders in the context of biology lessons. Pupils were taught under different instructional conditions, one of which combined specially designed lessons with a computer simulation, and their conceptual systems-thinking scores were compared across the conditions. Only the group that received both the special lessons and the simulation improved. The result is informative in two directions. It shows that systems thinking in this sense can be improved in sixth-grade pupils within ordinary schooling, in a study large enough and controlled enough to take seriously. It also shows that the improvement was not automatic: in this study, the conditions that did not combine the lessons with the simulation did not produce the gain. The design was quasi-experimental rather than randomised, and the outcome was a conceptual test rather than a measure of later behaviour, but among classroom studies it is one of the clearest.

A smaller study adds an important qualification. Evagorou et al. (2009) gave five 90-minute lessons using interactive simulations of a marsh ecosystem to children aged 11 and 12. Some systems-thinking skills improved. Feedback thinking did not. That is a notable result for a book about systems thinking, because feedback is the concept at its centre, and it suggests that the elements of the skill are not equally easy to teach. Recognising parts and interconnections may come readily; tracing how an effect loops back to its cause may need more than a few lessons, even with a simulation.

Mark for the third question: some real progress, from mostly weak designs. Children's systems thinking can be improved, and the best evidence suggests that simulation combined with structured teaching does more than either alone. Most studies are pre/post or case designs without comparison groups, outcomes are measured soon after teaching, and feedback thinking appears to be the hardest element to shift. The classroom evidence resembles the adult evidence on short courses that Chapter 5 described, where I graded the claim that brief training improves stock-flow reasoning C.

Do adults carry it out of the classroom?

The adult training studies described in Chapter 5 showed that short courses can improve people's reasoning about accumulation, including in the MIT study in which graduate students improved substantially while a minority still reasoned correlationally (Sterman 2010), and in a crash course reported with pre/post gains (Kainz and Ossimitz 2002). Those studies measured performance on tasks similar to the ones taught. The question here is the one that matters most to anyone commissioning training: does the skill transfer to decisions outside the classroom?

The best evidence comes from laboratory simulations that stand in for real decisions. Maani and Maharaj (2004) recorded participants thinking aloud as they managed a simulated business and analysed the protocols. The amount of systems thinking participants displayed did not, by itself, predict how well they performed. What distinguished the better performers was a pattern: they first sought to understand the structure of the situation before

acting, and their reasoning then followed a recognisable cyclical pattern. The sample was small and the task was a laboratory simulation, but the result is instructive. It suggests that what matters may be when and how systems thinking is used, rather than how much of it a person does.

Capelo and Dias (2009) tested a related idea in a management simulation built on a system dynamics model. Participants who reviewed a strategy map developed mental models closer to the true structure of the simulated business, and the closeness of their mental models in turn predicted their performance. Two of the study's three hypotheses were supported. This is the closest the literature comes to showing the chain the field assumes: a tool improves understanding of structure, and better understanding of structure improves decisions. It was shown in a simulation, though, over a short period, with the true structure known to the experimenters.

Beyond these laboratory proxies, the evidence stops. No randomised trial or longitudinal study was found showing that systems-thinking or system dynamics training improves decisions or outcomes in real organisations or policy settings. I'd grade the claim that systems-thinking training improves real-world decisions D: the support comes from laboratory simulations and protocol studies, one of which found that the amount of systems thinking did not predict performance, and the step from the laboratory to real decisions has not been tested. This is, in my judgment, the largest single gap between what the field promises and what it has shown, and it is the gap on which the practical case for most of the methods in Part III ultimately depends.

It would be wrong to suggest that the missing studies are easy. Real decisions are made infrequently, under varying conditions, with outcomes that arrive late and are shaped by much besides the decision-maker's reasoning. But feasible designs exist. Many organisations make the same kind of decision repeatedly, such as ordering stock, setting staffing levels or scheduling maintenance, and keep records of both the decisions and their consequences. A study could train some decision-makers in stock-flow reasoning and not others, perhaps staggering the training so that everyone eventually receives it, and compare the quality of their recurring decisions

over the following months against a benchmark such as the cost of shortages and surpluses. Such a design would not prove that systems thinking improves every kind of decision. It would show whether it improves one kind, in a real setting, which is more than the literature can currently claim.

Mark for the fourth question: untested. There are encouraging laboratory results about how structural understanding relates to performance in simulations. There is no evidence yet that training produces better decisions outside them.

The report card, read as a whole

Read together, the four marks tell a consistent story. The skill is defined in several ways that have not been reconciled. It is measured by instruments that are individually respectable but collectively fragmented, and whose overlap with general ability is unresolved. It can be improved in classrooms, most convincingly when simulation is combined with structured teaching, though feedback thinking resists and most designs are weak. And its transfer to real decisions has not been tested.

That profile has a specific implication for the leverage illusion. The promise that trained systems thinkers will see structure accurately and act on it well is a promise about the fourth mark, which is the one without evidence. The first three marks are the foundations that any test of the fourth would need. Without an agreed definition and convergent measures, it would not even be clear what a transfer study should measure. The field's most pressing educational task is therefore better assessment, which matters more at this stage than further courses, and the practical tool for this chapter is designed with that in mind.

It is also fair to note what the record does not show. It does not show that systems thinking cannot be taught; the Riess and Mischo results and the adult short-course studies suggest it can, in at least some of its elements. It does not show that training fails to transfer; nobody has run the study that would tell us. The record shows a field that has invested far more in teaching the skill than in finding out whether the teaching works where it matters.

An assessment blueprint that can tell the difference

The blueprint below is designed for anyone running a course, workshop series or training programme in systems thinking who wants to know what it achieved. It pairs a performance task with a self-report scale and a transfer task, adds a check on general reasoning, and times the measures so that they can separate different kinds of change. It is a design built from the gaps in the literature, not a validated package, and its value lies in the comparisons it makes possible.

1. **Choose a performance task tied to the course's goals.** Stock-flow problems in the style of the systems thinking inventory suit a course on accumulation and delay. A cognitive-mapping assessment of the kind Plate developed suits a course on causal structure. Use the Stave and Hopper taxonomy to state which levels of skill the course targets, and pick tasks that test those levels.
2. **Add an established self-report scale.** Choose one whose development has been published, such as the scale by Dolansky and colleagues in health care or one of the Stroink-group measures, instead of writing new items. The scale measures disposition and confidence, which are worth knowing about in their own right.
3. **Design a transfer task in a different domain and format.** If the course teaches with graphs about inventories, the transfer task should pose a structurally similar problem in words about, say, a hospital waiting list or a household budget. Chapter 2 described how strongly the presentation format affects performance, so a change of format is part of the test.
4. **Include a short reasoning test.** A brief measure of general reasoning allows a check on whether gains on the systems tasks simply track general ability, the question that Chapter 4 showed can change the conclusion.
5. **Time the measures.** Administer everything before the course, at its end and again three to six months later. The delayed measure distinguishes lasting learning from short-term familiarity with the task format.

6. **Find a comparison group if at all possible.** A waiting-list group, a cohort from the previous year or a similar course without the systems component turns a pre/post design into something closer to the quasi-experiments that give this chapter its best evidence.
7. **Record the relationships among the measures.** Report not only whether each score rose but whether the scores move together.

The value of the blueprint is in how the results are read. If the self-report scale rises and the performance task does not, participants have gained confidence rather than skill, which is worth knowing before anyone claims otherwise. If the performance task rises and the transfer task does not, the learning is tied to the taught format, which is a common and honest result. If all three rise but the gains track the reasoning test closely, the course may be selecting or rewarding general ability rather than teaching something distinct. If the performance and transfer gains hold at the delayed measure and exceed those of the comparison group, the course has evidence stronger than most of the literature reviewed in this chapter.

A worked illustration shows what this looks like in practice. Consider a hypothetical ten-week course in systems thinking for public health staff. Before the course, participants complete three stock-flow items about disease prevalence presented as graphs, the 20-item self-report scale, one transfer problem posed in words about a housing waiting list, and a ten-minute reasoning test. At the end of the course, and again four months later, they complete parallel versions of the same measures. A group of colleagues who will take the course the following year completes the same measures at the same times. Suppose, for the sake of the illustration, that self-report scores rise sharply at the end of the course and stay high, that the stock-flow items improve at the end but drift back halfway by four months, and that the transfer problem shows no change relative to the comparison group. The course team would then know three things it could not otherwise have known: that participants feel more capable, that their graph-based skill is real but fading, and that it has not yet carried over to a differently framed problem. Each finding suggests a specific change to the course, and none of them would be visible from a satisfaction survey.

The grade for what this blueprint can deliver depends on how much of it is used. Without a comparison group it yields pre/post evidence of the kind that earns a C. With a comparison group and a delayed measure, it can produce evidence stronger than much of what the field currently has. And if the transfer task were ever replaced by a measure of real decisions tracked over time, it would begin to answer the question this chapter has left open, whether systems thinking, once taught, changes what people do. Part IV turns from the toolkit to four attempts to use it on the largest scale, where that question was answered, or left unanswered, in public.

Part IV. Pulling the Big Levers

The chapters in this part leave the laboratory and the workshop behind and look at four occasions when people tried to move very large systems by working on their structure: the world economy and its resource base, the economy of one nation steered from an operations room, the health of whole populations, and the rules by which communities share what they cannot divide. In each case the question is the same. What lever was claimed, was it ever actually pulled, and what does the evidence say followed? Two of the four end with credit that is earned in full, one for Elinor Ostrom's commons and one for a carefully designed trial in Victoria.

Chapter 15. The Limits to Growth at Fifty: A Trial in Three Acts

Few arguments in the systems tradition have lasted as long, or been conducted with as much heat, as the argument over *The Limits to Growth*. The book appeared in 1972, and it is still being re-run, recalibrated and compared with new data half a century later. Its defenders say events have borne it out. Its critics say it was never built to be borne out by anything. Both sides have serious people on them, and both sides have been partly right.

That makes it a good case to try in the manner of a court. The prosecution in this trial is the economists' critique, which came fast and hard in the 1970s and has never been fully answered. The defence is the modellers' own account of what they were doing, together with the data comparisons that began to appear a generation later. The verdict, which I give in my own voice at the end, has to separate three questions that the public argument has usually run together: whether the model was well grounded when it was built, whether its main scenario has tracked the world since, and whether any of this tells us that the model's recommended lever would have worked.

This chapter owns several ideas that later chapters lean on: the world models themselves, the critique of their data, and the difference between validating a model and building confidence in it. It closes with a set of questions that can be put to any model, big or small, that someone asks you to trust.

Act I: the model and its makers

The story starts a year before the famous book. In 1971 Jay Forrester published *World Dynamics* with Wright-Allen Press, setting out a model called World2 that he had developed for the Club of Rome (Forrester 1971b). Chapter 8 follows Forrester's career and the method he founded; here the point is that World2 was the direct precursor of the model at the centre of this chapter. Forrester was arguing at the time that social systems behave counterintuitively, and a global model was the largest canvas on which that argument could be painted.

The following year a team led by Donella Meadows, with Dennis Meadows, Randers and Behrens, published *The Limits to Growth* with Universe Books in New York (Meadows et al. 1972). Its model, World3, linked the growth of population and industrial output to the physical limits on which that growth depends, through a set of feedback loops, and the book presented a set of scenarios. Under assumptions that amounted to business as usual, population and industrial output overshot their limits and then collapsed at some point within the twenty-first century. Other scenarios, with different assumptions about technology or about deliberate stabilisation, produced different paths.

One detail of framing matters more than any other for the trial that follows. The authors presented their runs as behaviour modes and explicitly denied that they were forecasts (Meadows et al. 1972). A behaviour mode is a qualitative shape, such as growth followed by overshoot and decline, and the claim attached to it is structural: given feedback loops of this kind, growth of this kind tends to end in this way unless the structure changes. A forecast is a claim about particular values at particular times. The distinction was sincere and methodologically defensible, yet it has been lost in almost every popular retelling. Critics attacked the book as a failed prophecy, and some admirers later praised it as a successful one. Both readings treat the standard run as a forecast, which is exactly what its authors said it was not.

It helps to see why a feedback model produces overshoot at all, since the shape carries the whole argument. Chapter 1 introduced the idea that a stock changes only through its flows. Suppose a growing economy draws on a

stock of some resource, and suppose that the signal telling the economy the stock is running low arrives only after a delay, because exploration, prices and investment decisions all take time to register depletion. While the signal is in transit, growth continues at full speed. By the time the economy responds, it has already drawn the stock down further than it would have if the signal had arrived at once, and its own capital, built for a larger flow, keeps pulling on the stock after the decision to slow has been taken. The system overshoots the level it could have sustained. If the drawdown also damages the stock's capacity to regenerate, or if the pollution that accompanies growth accumulates with its own delay, the overshoot can turn into decline. None of this needs precise numbers. It follows from growth, a finite or slowly renewing stock, and a delayed response.

The modellers did not stop at defending their work. A decade later Donella Meadows, Richardson and Bruckmann published *Groping in the Dark*, a reflective comparison of World3 with the rival global models that had appeared in the meantime (Meadows, Richardson and Bruckmann 1982). The book is frank about the limits of the whole enterprise, including the limits of the modellers' own work. That frankness is worth recording, because the rest of this chapter will sometimes sound as if the modellers were defending an indefensible position. They were more self-critical than most of their critics credited, and the title of the 1982 book says as much.

The last piece of Act I is the updates. In 2004 Donella Meadows, Randers and Dennis Meadows published a thirty-year update built on a revised model, World3-03, and argued that humanity was already in overshoot (Meadows, Randers and Meadows 2004). The update is a statement of the authors' judgment, supported by the revised scenarios. It is best read alongside the independent data comparisons in Act III, which were made by people outside the original team.

Why should a practitioner care about how World3 was built, rather than simply what it said? The answer is that the lever the book recommended was a structural one. If the loops linking growth to the limits it depends on are the ones the model describes, then the place to act is on growth itself, early, before the delays in the system make correction impossible. That is a

leverage claim in the full sense this book examines: see the structure, and the structure tells you where to push. Whether the claim deserves belief depends on whether the structure was right, and that is what the prosecution set out to challenge.

A hypothetical makes the logic of early action concrete. The numbers here are invented and belong to no model. Suppose an activity grows at 7 per cent a year, so that it doubles roughly every ten years, and suppose that a decision to change course takes 20 years to take full effect once made, because the equipment, buildings and habits of the growing system last that long. A decision taken today therefore lands on an activity about four times its present size. A decision postponed by ten years lands on one about eight times its present size. The lever, on this reasoning, loses force with every year of delay, and that is why the book's authors argued for acting before the limits were visible. The argument is sound as arithmetic. Whether it applies to the real world depends entirely on whether the growth, the limits and the delays are as the model assumes, which is the question Act II raises.

Act II: the economists' brief

The first and most quoted attack came from William Nordhaus. His 1973 review in *The Economic Journal* carried the title "World Dynamics: Measurement Without Data", and the title summarised the charge (Nordhaus 1973). Nordhaus argued that Forrester's world model rested on relationships that had never been estimated from empirical data, and that its conclusions depended on assumptions nobody had tested, above all assumptions about technology and about substitution between resources. Nordhaus's target was Forrester's *World Dynamics* rather than World3, but the charge travelled, because the two models shared a method and a family resemblance.

Robert Solow's 1973 essay "Is the End of the World at Hand?" made a related case in plainer language (Solow 1973). A model of the world economy that leaves out prices, substitution and technical change, Solow argued, leaves out the very mechanisms by which economies respond to scarcity. When a resource becomes scarce its price rises, people use less of it and look for

alternatives, and engineers find ways of doing more with less. A model without those responses will find limits sooner than an economy with them, because it has removed the feedback loops that push the limits back.

A hypothetical example shows the force of Solow's objection. Imagine a model in which an economy uses a fixed amount of a metal for each unit of output, and the known reserves of that metal are fixed. Run it forward with output growing, and the metal runs out on a date you can compute with a calculator. Now add three responses that real economies display. As the metal grows scarce its price rises, and users economise on it. Higher prices make lower-grade deposits worth mining, so the effective reserve grows. Engineers substitute other materials in some uses. Each response is a balancing loop, and together they push the exhaustion date back, perhaps indefinitely. Whether these loops are strong enough in any particular case is an empirical question. Solow's point was that a model which omits them has answered the question by assumption.

The most detailed critique came from the Science Policy Research Unit at the University of Sussex. *Models of Doom*, edited by Freeman, Jahoda and colleagues and published by Universe Books in 1973, went through World3's sub-models and data one by one (Freeman, Jahoda et al. 1973). The volume included a reply from the MIT team, which is a courtesy worth noting: the critics printed the defence alongside the charge.

A sub-model-by-sub-model critique matters because of how feedback models behave. In a model with many interlocking loops, a modest error in one relationship can be amplified or damped by the others, and it is not obvious in advance which. The only way to find out is to take the model apart, examine each relationship against whatever evidence exists, and see how much the overall behaviour depends on the weakest pieces. That is what the Sussex group attempted. Their volume was the first large-scale exercise of the kind for a global model, and it established a standard that later critics and defenders alike have had to meet.

Forrester, Low and Mass answered Nordhaus directly in *Policy Sciences* in 1974 (Forrester, Low and Mass 1974). They defended the structure of *World Dynamics* and, more interestingly, defended system dynamics' approach to

evidence. The system dynamics position, as it developed over the following decades, was that a model's structure can be grounded in many kinds of knowledge, including operating experience, physical constraints and qualitative descriptions of decision-making, and that econometric estimation of each parameter is neither necessary nor always possible. Barlas later formalised this view, setting out tests of structure that come before tests of behaviour, while also observing that the field had put little explicit effort into validation (Barlas 1996).

Nearly two decades after his first review, Nordhaus returned to the argument. His 1992 paper in the *Brookings Papers on Economic Activity*, "Lethal Model 2: The Limits to Growth Revisited", brought growth theory and data to bear on the limits question and concluded that resource and environmental constraints should prove modest over the following 50 years, provided institutions and incentives were sound (Nordhaus 1992). The conditional in Nordhaus's conclusion matters. He did not argue that limits were imaginary. He argued that the economy's adaptive mechanisms would handle them if the institutions that carry price signals and incentives worked.

How should the prosecution's case be graded? The claim at issue is that Nordhaus was right that *World Dynamics* lacked empirical grounding, and I'd grade it B. The core factual points, that the parameters were not estimated econometrically and that the model had no price mechanism, are accepted even in sympathetic later work. A review of the criticisms of system dynamics written from inside the field judged several of them still pertinent, the role of historical data among them (Featherston and Doolan 2012), and the modellers' own reflections in *Groping in the Dark* are candid about limits. The grade stops at B because the further question, whether that lack of grounding disqualifies the model's conclusions, is a dispute about values and method that evidence alone cannot settle.

That value dispute deserves a sentence of its own, because it explains why the argument never ended. The economists and the modellers disagreed about time horizons, about how much faith to place in adaptive mechanisms that had worked in the past, and about which kind of error was worse:

acting against limits that turned out to be distant, or failing to act against limits that turned out to be near. Those are judgments about risk and about the future. A model can inform them, but it cannot make them.

Act III: the data come in

For a generation the argument was conducted mainly on grounds of method. Then, in 2008, Turner, working at CSIRO, did something that neither side had done systematically: he compared the World3 scenarios with 30 years of observed data (Turner 2008).

Because the details of Turner's study are often misreported, it is worth setting out what he actually did. Turner took observed global data for 1970 to 2000 on the variables World3 reports and compared them with the output of three of the original scenarios: the standard run, the comprehensive-technology run and the stabilised-world run. He measured fit with a normalised root-mean-square deviation. For the standard run, every variable except the death rate came in with a normalised deviation below 20 per cent. The data did not compare well with the comprehensive-technology or stabilised-world scenarios (Turner 2008).

A word on what the fit measure means may help readers who have not met it. A root-mean-square deviation takes the gap between the model's value and the observed value at each point in time, squares it so that misses in either direction count, averages the squares and takes the square root. Normalising expresses the result relative to the size of the observed values, so that variables measured in different units can be compared. A normalised deviation below 20 per cent means, roughly, that the model's path and the observed path stayed within a band of that relative width over the period. That is a respectable fit for a model built decades earlier. It is also a coarse measure, and it says nothing about whether the model got the right answer for the right reasons.

The comparison also discriminated between scenarios, and that deserves credit. The comprehensive-technology and stabilised-world scenarios had begun to depart from the standard run within the observed period, because their assumptions about policy and technology changed the system's

behaviour early. The data could therefore tell them apart, and the data sided with the standard run. What the data could not do was tell the standard run apart from rival models of different structure that would have tracked the same 30 years equally well. The first kind of discrimination is real evidence. The second is where the arguments of the cross-examination below come in.

Turner's own words, as given in the CSIRO working-paper version of the study, are that the observed data "compares favorably with key features of a business-as-usual scenario" (Turner 2008). He went further than many readers assume. In the working-paper version he writes that the good comparison "provides a degree of validation of the World3 model", and that it "lends support to the conclusion from the LtG that the global system is on an unsustainable trajectory" (Turner 2008). He also flagged the limits of that validation, describing it as constrained by the way model output is reported, and he noted that the comparison was "as much a test of the scenarios as it is of the model" (Turner 2008). Turner, then, claimed a real but limited degree of validation and was careful to say where its limits lay. Any account that has him disclaiming validation altogether gets him wrong.

Later comparisons continued in the same direction. Herrington updated the data comparison and found that the two scenarios closest to observed data implied a halt in the growth of welfare, food and industrial output within about a decade (Herrington 2021). Nebel and colleagues recalibrated the World3-03 parameters against empirical data and found that the recalibrated run still produced overshoot and collapse, with the peaks somewhat higher and a few years later (Nebel et al. 2024). Bardi's history of the debate, sympathetic to the modellers, placed these comparisons within the long story of the book's reception (Bardi 2011).

The recalibration deserves a closer look, because it answers a different question from the data comparisons. Turner and Herrington asked whether the old scenarios had tracked the data. Nebel and colleagues asked whether the model's characteristic behaviour depended on the particular parameter values chosen in the 1970s. If overshoot had vanished once the parameters were fitted to modern data, that would have suggested the famous shape was

an artefact of guesswork. It did not vanish. That is a meaningful result of the kind modellers call a robustness test: it indicates that the behaviour mode comes from the model's loops rather than from a lucky choice of numbers. What a recalibration cannot do is tell you whether those loops are the right ones. A model with the wrong structure, fitted carefully to data, will reproduce its own wrong behaviour just as faithfully.

This is the defence at its strongest, and it deserves a fair hearing. A model built in the early 1970s, attacked at the time for lacking data, produced a baseline scenario whose main aggregates stayed close to observed values for three decades. The rival scenarios in the same model did not. Recalibrating the parameters against data did not make the overshoot go away. None of this was guaranteed in advance, and a model with badly wrong structure could easily have failed the comparison.

Now the cross-examination. Castro argued in 2012 that claims of imminent collapse based on World3 data-fits are premature, given the structural uncertainty in the simulation (Castro 2012). The argument has deeper roots. In a much-cited 1994 paper in *Science*, Oreskes, Shrader-Frechette and Belitz argued that numerical models of open natural systems cannot be verified or validated in any strict sense (Oreskes, Shrader-Frechette and Belitz 1994). A model can be evaluated, and it can be partially confirmed by agreement with observation, but agreement never proves that the model's structure is correct, because other structures could have produced the same agreement.

A simple illustration shows why this bites. The numbers that follow are invented for the purpose and describe no real model. Imagine two models of a growing quantity, each fitted to 30 years of data in which the quantity roughly doubles every 15 years. Model A contains a limit that erodes as the quantity grows, so that after year 40 growth reverses and the quantity falls. Model B contains a limit that simply slows growth, so that after year 40 the quantity levels off and stays high. For the first 30 years the two models can be made almost indistinguishable, because both are dominated by the same growth loop and neither limit has yet begun to bind. Their fit to the data might be equally good. Yet they say opposite things about year 50. Data from

years one to 30 cannot choose between them, however closely each tracks the record. The fit tests the stretch of history the models share, and says little about the future that separates them.

That is the situation of the World3 scenarios. They diverge mainly after the period of observed data, and so the observed data cannot yet discriminate between them in the way that matters most. Turner himself pointed at a version of this when he wrote that his comparison tested the scenarios as much as the model.

The cross-examination leads to two grades. The first concerns the claim that World3's standard run has tracked observed data since 1972. I'd grade it C. The evidence is real and has been reproduced by more than one analyst, but it consists of aggregate comparisons over a period in which the competing scenarios had not yet separated, and tracking a handful of aggregate indicators over decades is not a test of the model's structure. The second concerns a broader claim, that system dynamics models can be validated in a meaningful sense. This also gets a C, for a different reason: the disagreement is philosophical rather than empirical. Barlas's structure and behaviour tests on one side and Oreskes and colleagues on the other are arguing about what validation can mean, and no experiment will resolve that. Saltelli and colleagues' manifesto on responsible modelling, with its warnings about hidden assumptions, hubris and unknowns, adds a modern voice to the cautious side (Saltelli et al. 2020).

The practical upshot, and it is one that runs through the rest of this book, is to speak of building confidence in a model rather than validating it. Confidence grows as a model passes more tests of different kinds: tests of whether its structure matches what is known about the real system, tests of whether its behaviour matches history, tests of whether it behaves sensibly under extreme conditions, and tests of whether its policy conclusions survive changes in uncertain assumptions. Each passed test adds something. None of them, alone or together, turns a model into a proven account of the world.

Judgment on four counts

I will give the verdict in four parts, because the three questions I separated at the start have separate answers, and there is a fourth point about levers that the trial format tends to hide.

On grounding, the prosecution largely wins. Nordhaus was right on the facts: the world models of the early 1970s were built with parameters that had not been estimated from data and without the price responses that economists regard as central to how economies meet scarcity. That was a real weakness, acknowledged in time by people who were sympathetic to the modellers. Where the prosecution overreached was in treating the weakness as decisive. A model with poorly estimated parameters can still capture the qualitative shape of a system's behaviour, and the dispute about whether World3 did so was a dispute about values, method and time horizons as much as about models.

On tracking, the defence has a genuine point, and I think it should be granted without grudging. The standard run's aggregates have stayed close to several observed series, and they have done so better than the model's other scenarios. That is interesting, and it is more than many models of comparable age could claim. But it is not a test of structure. The fit covers a period in which rival structures would behave alike, and the scenarios part company mainly in the years that have not yet been observed. Turner's own phrase for what his comparison provides, a degree of validation, is about as far as the evidence goes, and the more cautious reading that aggregate tracking is not a structural test, which draws on Oreskes and on Castro, is the one this book adopts.

On the forecast question, a caution to both camps. The authors said their runs were behaviour modes, not forecasts. Admirers who now treat the standard run as a vindicated forecast are applying a standard the authors disowned, and critics who once treated it as a failed forecast were doing the same.

The fourth point concerns the lever. World3's recommended intervention was to slow and eventually stabilise growth, deliberately and early. That lever was never pulled. The world did not adopt the stabilised-world

policies, and so the history since 1972 is, at most, a test of what the model said would happen if nobody acted on it. The case can tell us something about the model's baseline. It can tell us nothing about whether the structural lever it identified would have delivered the stabilised world it promised, because nobody tried. In the vocabulary of this book, *The Limits to Growth* is a case in which the lever was claimed, clearly and with real intellectual seriousness, and then left untested by history.

This distinction has a practical edge. When a model's baseline appears to have come true, there is a strong temptation to transfer the credibility of the baseline to the model's recommended policy, as if a model that saw the problem coming must also know the cure. The two claims rest on different parts of the model. The baseline depends on the loops that operate when nobody intervenes. The policy claim depends on how the system would respond to a deliberate change, which involves the responses of people, markets and institutions that the baseline may never have exercised. A good baseline is a reason to take the model seriously, but on its own it tells us nothing about the lever, and the leverage illusion is at its most persuasive when this difference is blurred.

The successors of World3 carry the same structure of argument in a new form. In 2009 Rockström and colleagues proposed nine planetary boundaries that together define a safe operating space for humanity, and estimated that several had already been crossed (Rockström et al. 2009). A 2023 update by Richardson and colleagues concluded that six of the nine had been transgressed (Richardson et al. 2023). Alongside the boundaries runs the literature on tipping points. Lenton and colleagues defined tipping elements in the climate system in 2008, drawing up a shortlist from the literature, a workshop and expert elicitation (Lenton et al. 2008). A 2019 comment in *Nature* by Lenton and colleagues warned that climate tipping points may be closer than previously thought (Lenton et al. 2019), and a 2022 synthesis led by Armstrong McKay revised the list of tipping elements and their temperature thresholds, concluding that warming beyond 1.5°C could trigger several of them (Armstrong McKay et al. 2022).

These frameworks are more carefully bounded than World3. They do not try to model the world economy as a single feedback system, and they rest on a much larger body of physical evidence. But they share World3's central logical move: they identify thresholds in a system whose most important behaviour lies in a future that cannot be observed in advance, and they recommend acting before the evidence of crossing becomes unambiguous. The same questions about confidence, structure and values apply. Chapter 18 takes up the related question of whether approaching thresholds can be detected from data in advance, which is where early-warning signals come in, and finds the answer less encouraging than the idea suggests.

Questions to put to any model

John Sterman, whose experiments anchor Part I of this book, wrote a short piece in 1991 called "A skeptic's guide to computer models" (Sterman 1991). It was a modeller's own checklist for doubting models, organised around their boundaries, their assumptions, their validation and the purpose they were built for. The questions below are built around those themes and adapted to the lessons of the World3 trial. They are not Sterman's exact wording, and they work as well on a spreadsheet forecast or a consultant's causal diagram as on a global simulation. The underlying theory is sound, but whether asking them improves decisions has not been tested, so treat them as a discipline of attention rather than a validated instrument.

1. **What is the model for?** A model built to explore behaviour modes should not be judged, or sold, as a forecasting tool. Ask the builders to state the purpose in one sentence, and check whether the conclusions you are being offered fall inside it.
2. **What is inside the boundary, and what has been left out?** The economists' objection to the world models was about omitted feedback: prices, substitution, technical change. For any model, list the responses the real system can make that the model cannot. Each omission is a place where the model may find limits too early or too late.

3. **Where did the numbers come from?** For each important parameter, ask whether it was estimated from data, taken from another study, elicited from experts or chosen to make the model behave plausibly. None of these is illegitimate, but they deserve different amounts of trust, and you are entitled to know which is which.
4. **What tests has it passed, and of what kind?** Ask separately about tests of structure (does each relationship correspond to something known about the real system?), tests of behaviour (does it reproduce history?) and tests of robustness (does the policy conclusion survive when uncertain assumptions change?). A model that has passed only the second kind has shown that it can fit a curve.
5. **Would a different structure fit the same history?** This is the Oreskes question and the lesson of the two invented models above. If the builders cannot name a rival structure and explain why the data favour theirs, the fit is weaker evidence than it looks.
6. **Where do the scenarios diverge?** If the scenarios separate only in the future, historical fit cannot choose between them. Find the date at which the paths part and ask what evidence, available before that date, would tell you which one you are on.
7. **What is the recommended lever, and has anyone pulled it?** Separate the model's forecast of what happens without intervention from its claim about what an intervention would do. The second claim needs its own evidence, and in the case of World3 there was none to be had.
8. **Who bears the cost of each kind of error?** Much of the fifty-year argument was about values and time horizons presented as a dispute about models. Say openly which error you fear more, acting too soon or acting too late, and why, so that the disagreement is conducted where it actually lies.

To see how the questions work in practice, consider a hypothetical. It is invented for illustration and describes no real authority. A regional water authority is shown a simulation that projects demand exceeding supply within 15 years and recommends an immediate programme of restrictions. Asked what the model is for, its builders say it was built to explore the effect

of population growth on demand, which fits the question. Asked what lies outside the boundary, they concede that the model holds water prices constant and has no loop by which households or farms respond to tighter supply by using less, which is the Solow objection in a new setting. Asked where the numbers came from, they explain that demand per household was taken from a single survey and the rate of population growth from a planning forecast. Asked what tests the model has passed, they show a close fit to 20 years of demand data, but no tests of structure and no test of whether the recommendation survives a different price assumption. Asked whether a different structure would fit the same history, they have not checked. None of this shows the model is wrong. It shows that the authority is being asked to act on a behaviour-mode model as if it were a forecast, and it tells the authority which two or three additional tests would most change its confidence.

Asked seriously, these questions rarely end with a model being thrown out. More often they end with a clearer sense of what the model can and cannot tell you, which is the most any model can offer, and more than *The Limits to Growth* was usually granted by either its admirers or its critics.

Chapter 16. Cybersyn: Viable Systems and the Dream of Real-Time Control

Of the four attempts to pull a big lever that this part of the book examines, one comes closest to the literal picture the phrase suggests. Between 1971 and 1973, in Salvador Allende's Chile, the British cybernetician Stafford Beer led a project to connect the country's economy to a centre from which it could be watched and steered as it changed. The project was called Cybersyn. Its intellectual foundation was a model of organisation that Beer had spent years developing, and its political setting was a socialist government trying to run a national economy without abandoning its commitment to the participation of workers. It operated only in part, for less than two years, and it ended with the coup of 1973 (Medina 2006; Medina 2011).

Cybersyn has become something of a legend, and legends attract detail. This chapter keeps to what the archival histories and the sources behind this book support. In particular, I have left out the physical descriptions of the project's operations room that decorate many popular accounts, because I have not been able to confirm them from Medina's text, which is the authoritative reconstruction. What remains is less cinematic and, I think, more instructive. It is the story of a serious theory of control applied to the largest target anyone has tried, and of a historical record that cannot, by its nature, tell us whether it worked.

A model of the firm borrowed from the nervous system

Beer set out the first full statement of what he called the Viable System Model in *Brain of the Firm* in 1972 (Beer 1972). As the title suggests, the model rested on an analogy with the nervous system. Beer asked what any

organisation must contain if it is to survive in a changing environment, and he answered by describing five interacting functions, which he numbered System 1 to System 5. He later derived the same structure from first principles of what he called variety engineering in *The Heart of Enterprise* (Beer 1979) and wrote a practitioner's manual for using it to diagnose organisations, *Diagnosing the System for Organizations* (Beer 1985).

In outline, the five systems divide the work of staying viable as follows. System 1 is the set of operational units that do the organisation's primary work: the plants, teams or services that produce whatever the organisation exists to produce. System 2 coordinates those units, damping the oscillations and conflicts that arise when semi-independent units share resources or depend on one another. System 3 manages the operational whole in the present, allocating resources among the units and holding them to account. System 4 looks outward and forward, at the environment and at the future, asking what is changing and how the organisation must adapt. System 5 sets policy and holds the organisation's identity, and one of its main tasks is to balance the pull of System 3, which wants stability and efficiency now, against the pull of System 4, which wants change for the sake of later. A further feature of the model is recursion: each operational unit in System 1 is itself a viable system, with its own Systems 1 to 5, and the organisation as a whole is in turn one unit within a larger viable system.

Two things about this model are worth noticing before going further. The first is that it is a functional description. It does not say who should perform each function or how the organisation chart should look. A small team may have all five functions performed by the same three people, and a large firm may spread them across whole departments. The diagnosis Beer proposed asks whether each function is being performed adequately and whether the channels between them carry enough information, and the answer can be yes in organisations that look very different on paper.

The second is that the model gives a great deal of weight to the autonomy of the operational units. System 3 does not run System 1 in detail. It sets the conditions within which the units run themselves, and it intervenes only where their autonomy threatens the cohesion of the whole. This emphasis

on autonomy matters for how Cybersyn should be read. A model that could be caricatured as a machine for central control was, in its author's intention, a theory of how much freedom the parts of a system can have while the whole stays coherent.

The Viable System Model attracted a community of practitioners. Espejo and Harnden's 1989 collection gathered interpretations and applications, mostly in the form of practitioners' case accounts (Espejo and Harnden 1989), and later writers have argued that the model puts the principles of complexity science into practice, including in support of self-organising governance for sustainability (Espinosa, Harnden and Walker 2008). Those later applications are illustrative rather than evaluated. I know of no controlled comparison in the bibliography behind this book showing that organisations diagnosed and redesigned with the model outperform comparable organisations that were not.

Engineering variety

Chapter 6 introduced Ashby's law of requisite variety, the principle that a regulator can hold a system steady only if it commands as much variety, as many distinct responses, as the disturbances it must counter. Beer took that law as the foundation of management. Any manager faces far more variety in the operations below and the environment outside than any single mind or office can match. The manager's problem, on Beer's account, is to engineer variety: to attenuate the variety arriving from below and outside, so that only what matters reaches the top, and to amplify the variety of the manager's own responses, so that a small number of decisions can shape a large number of outcomes.

A hypothetical example makes the idea concrete. Consider an invented regional bus company with 40 routes. On any given morning each route can be early, on time, late or cancelled, and each delay can have a dozen causes. No operations manager can attend to every combination. The company attenuates this variety by setting routes and timetables so that most mornings produce no news, by delegating most problems to depot supervisors who handle them without reporting upward, and by defining a small

number of exceptions that must be escalated. It amplifies the manager's variety by giving supervisors standing authority to act, by building rules that respond automatically to common disruptions, and by keeping spare buses and drivers in reserve. If the attenuators filter out something important, or the amplifiers are too weak to respond to what gets through, the company loses control even if everyone in it is competent, because the variety balance has failed.

This is a genuinely useful way of seeing an organisation, and it has a respectable theoretical basis. Conant and Ashby's good regulator theorem, which Chapter 6 also discussed, adds that under broad conditions an effective and simple regulator must be a model of the system it regulates (Conant and Ashby 1970). Read together, the two results suggest a design ambition. If a government could build a model of its economy, supplied with timely information and able to filter out the routine and amplify the response to the exceptional, it might regulate that economy in something close to real time.

That ambition is where the leverage illusion can enter. The mathematics of requisite variety and good regulation are statements about idealised regulators and systems. They say what is necessary for control. They do not say that any particular organisational design achieves it, and they say nothing about how people inside the system will respond to being modelled and steered. Chapter 6 traced the gap between cybernetics as mathematics and cybernetics as metaphor. Cybersyn was the attempt to close that gap at the scale of a nation.

There is a further subtlety in the idea of attenuation that matters for any attempt to regulate from a centre. Every attenuator embodies a judgment about what matters. When the bus company in the example decides which disruptions must be escalated, it is deciding which problems its managers will ever hear about, and by implication which ones they will never see. In a bus company the judgment is fairly safe, because the purpose of the organisation is clear and the consequences of a missed signal are usually small. In a national economy the choice of indicators is a choice about which outcomes the state will attend to and which it will leave to others, or to

nobody. The theory of variety engineering says that such choices must be made. It cannot say which choices are right, because that depends on purposes the theory does not supply.

Chile, 1971 to 1973

The archival history of Cybersyn is Medina's work: first a long article in the *Journal of Latin American Studies* in 2006, titled "Designing Freedom, Regulating a Nation", and then a book, *Cybernetic Revolutionaries*, published by MIT Press in 2011 (Medina 2006; Medina 2011). Beer published his own account of the Chilean project in a collection of his papers, *Platform for Change* (Beer 1975). Pickering's history of British cybernetics places Beer among a group of thinkers, including Ashby, whose science was, in Pickering's reading, performative rather than representational: more concerned with building systems that adapt in action than with producing accurate pictures of the world (Pickering 2010).

The most important finding of Medina's research, for the purposes of this book, is that Cybersyn's politics were built into its technical design. The Allende government wanted two things that pulled in different directions. It wanted the capacity to coordinate a national economy centrally, and it wanted the workers in that economy to participate in running it. Medina shows that the tension shaped the network itself, instead of waiting for politicians to resolve it after the engineers had finished: what information would flow, to whom, and how much autonomy the local units would retain (Medina 2006). Medina's title catches the double aim in a phrase. The project was meant to design freedom and to regulate a nation at the same time.

That tension maps directly onto the Viable System Model. The model's emphasis on the autonomy of operational units gave Beer a theoretical reason to keep decisions local wherever possible and to send upward only what the centre genuinely needed. The government's need to coordinate an economy under severe political and economic pressure created a practical

reason to want more information at the centre and more intervention from it. A design that could serve both aims was an intellectual achievement. Whether it could hold under the pressure of events was never fully tested.

It was not fully tested because the system never ran in full. Medina's book documents limited operational use (Medina 2011). The project operated in part, for less than two years, and the coup of 1973 ended it along with the government that had commissioned it. Everything that might have told us whether a cybernetic model of a national economy could deliver better regulation than the alternatives, a period of full operation, comparison with the parts of the economy it did not reach, a record of decisions taken with and without it, was cut off before it could accumulate.

What a two-year experiment can show

It is worth being precise about what kind of evidence a historical case like Cybersyn provides, because the popular argument about it often slips between different questions.

A historical case can show that something was attempted, and how. Medina's research shows in detail that a team working with Beer designed and partly built a system intended to bring cybernetic regulation to a national economy, and it shows how political commitments shaped the technical choices. That is solid historical knowledge, and it matters for anyone who thinks the design of information systems can be separated from politics.

A historical case can also show what was possible under particular conditions: that the design problems could be worked through, that some parts of the system could be made to operate, that people could be persuaded to use it. Here the evidence is thinner, because partial operation for a short period under crisis conditions tells us about the early stages of a project rather than its mature performance.

What a single, interrupted case cannot show is whether the approach worked, in the sense of producing better outcomes than would otherwise have occurred. That question needs a comparison: the same economy without the system, other economies with and without similar systems, or

at least a long enough run to see whether the regulated parts behaved differently from the unregulated ones. None of that exists for Cybersyn. The claim sometimes made on its behalf, that it showed cybernetic real-time governance works, therefore gets a D, because there is no outcome evidence in either direction; nothing in the record counts against the project either. The case is a rich piece of history and a demonstration of nothing.

It helps to imagine what an evaluation would have required, because the same requirements apply to any present-day attempt to steer a large system through a real-time information network. The following is a hypothetical design, not a description of anything that happened. One would want the system introduced to some enterprises or sectors before others, so that those not yet connected could serve as a comparison. One would want measures of the outcomes the system was meant to improve, such as the speed with which production problems were noticed and resolved, recorded in connected and unconnected units alike before and after the change. One would want a record of the decisions the centre took on the basis of the system's signals, and of what happened afterwards, so that the quality of those decisions could be judged. And one would want enough time for the system to settle past its teething problems. None of these conditions was met, and under the circumstances of Chile between 1971 and 1973 it would have been remarkable if any had been. Nobody should fault the project for lacking them, but without them no verdict on its effectiveness is possible.

The same reasoning cuts against Cybersyn's detractors. Those who treat it as proof that cybernetic planning was doomed are reading an outcome into a case that was ended by a coup before any outcome could be observed. The project was stopped before it could succeed or fail on its own terms. Neither its admirers nor its critics can claim the evidence of results, and a fair reader has to hold both judgments in suspension.

Revival and reappraisal

Cybersyn has had a lively afterlife. In 2014 Morozov published an essay in *The New Yorker*, "The Planning Machine", that linked the project to present-day algorithmic governance and the politics of big data (Morozov 2014).

The essay brought Cybersyn to a large new audience. It drew heavily on Medina's research, and it is best read as a sign of how Cybersyn is received today rather than as independent evidence about what happened in Chile.

The revival says something about the appeal of the leverage illusion. A project that promised a small team, a model and a flow of information could steer an entire economy is an irresistible story in an era when data flows are vastly larger and cheaper than anything available in 1972. The temptation is to conclude that Cybersyn was an idea ahead of its technology, and that what it lacked in computing power the present can supply. That conclusion skips the questions that the history leaves open: whether the model was adequate to the economy, whether the people being regulated would have accepted the regulation, and whether the tension between central coordination and local autonomy could have been held once the system was fully running.

The same questions arise, on a smaller scale, whenever an organisation installs a dashboard and expects it to deliver control. More data does little for the variety problem, which lies in the mismatch between the variety of what happens and the variety of what a regulator can respond to. A shortage of information is only a small part of that mismatch. A centre that receives more indicators must still decide which to act on, and each additional indicator is one more thing competing for limited attention. The judgments embedded in attenuation, about what to watch and what to ignore, remain where they were. The revival's enthusiasm for Cybersyn as a precursor of data-driven government tends to pass over this. The history suggests that the hard part of real-time control lay in the design of the filters and the politics of who set them, and those are not problems that cheaper computing solves.

The critical tradition within systems thinking raised some of those questions early. Ulrich published "A critique of pure cybernetic reason: the Chilean experience with cybernetics" in 1981 (Ulrich 1981), a critical appraisal of the project from the standpoint of critical systems thinking; I cite it here for its existence and its angle, since its detailed argument has not been examined for this book. Jackson's 1988 appreciation of Beer's viable-system viewpoint assessed both its value and its limits as a guide to

managerial practice, again from a critical-systems standpoint (Jackson 1988). The general line of this critique, as Chapter 9 describes for the critical tradition as a whole, is that a model of regulation carries implicit judgments about whose purposes the system serves and who draws its boundaries. A regulator designed to keep a system viable must be told what viability means, and that is a political choice presented as a technical one.

Freedom and control in the same design

The strongest general argument against projects like Cybersyn comes from outside the systems tradition. Scott's *Seeing Like a State* examined a series of large schemes to improve the human condition, drawn up by governments confident that they could see and order society from above (Scott 1998). Scott argued that such high-modernist planning fails because it erases local, practical knowledge, which he called mētis: the know-how of farmers, workers and residents that cannot be written into a plan because it lives in practice and responds to local conditions the planner never sees. A state that simplifies society to make it legible and controllable destroys some of what made society work.

Scott's argument is a powerful counterweight to control-oriented systems ambitions, and it applies to Cybersyn in obvious ways. An information system that sends selected indicators to a centre necessarily simplifies. Whatever the indicators leave out is invisible to the regulator, and a regulator acting on what it can see may override the local knowledge that would have handled what it cannot.

Yet the fairest reading of Cybersyn does not place it simply on the side of the high modernists. Medina's evidence that worker participation was designed into the system, and Beer's own model with its insistence on the autonomy of operational units, show that the project's designers were aware of the problem Scott later described and tried to build an answer into the architecture. Pickering's reading of British cybernetics as performative points the same way. Beer's instinct was to build systems that adapt and learn in operation, rather than to produce a master plan and impose it.

Whether that answer would have survived full operation, political pressure and the natural pull of any centre toward more control is exactly what the history cannot tell us.

Chapter 18 describes a different answer to the same problem, arrived at by a very different route. Elinor Ostrom's work on the commons found that resources shared by many users are often governed well by arrangements with many centres of decision rather than one, an idea she developed as polycentric governance (Ostrom 2010), and she warned against blueprint solutions applied uniformly to problems that differ (Ostrom, Janssen and Anderies 2007). Her evidence came from comparing many real institutions over long periods, which is exactly the kind of evidence Cybersyn never had the chance to generate. The contrast is instructive. Beer reasoned from a theory of viability toward a design; Ostrom reasoned from observed institutions toward principles. Both arrived at an emphasis on local autonomy within a coordinating frame, and only one of them was able to test the claim across cases.

I think the right conclusion is a narrow one. Cybersyn shows that a cybernetic theory of organisation can be turned into a serious design for national regulation, and that the design will carry political choices whether its authors acknowledge them or not. It does not show that the design worked, and it does not show that it would have failed. As a lever, it was built, partly connected, and then removed before anyone could see what pulling it would do.

A five-system check on a team you know

The Viable System Model's lasting value for a practitioner is as a diagnostic lens, the use Beer's 1985 manual was written to support. The version below is simplified and is meant for a single team or small unit that you know well. The evidence for the model as a way of improving organisational performance is weak, a D, because the published applications are practitioners' case accounts without comparisons. Treat the check as a structured way of noticing gaps, not as a measurement.

To show how it works, the following walk-through uses a hypothetical team, a six-person analytics group serving several departments in a mid-sized organisation. It is a composite invented for illustration.

1. **System 1: what does the team actually produce, and who does it?** List the primary outputs and the people or sub-teams who produce them. In the analytics group, the outputs are regular reports, one-off analyses and a small number of dashboards, with two people on each. Check whether each unit has enough autonomy and resources to do its work without constant escalation.
2. **System 2: what stops the units colliding?** Look for the mechanisms that coordinate shared resources and schedules: a shared calendar, a queue, standard formats, a weekly stand-up. In the example, the report writers and the dashboard builders draw on the same data engineer, and nothing schedules that person's time. The symptom is recurring conflict and oscillation between feast and famine, which is what an absent coordination function tends to produce.
3. **System 3: who allocates resources and checks performance in the present?** Identify how work is prioritised and how the team knows whether it is delivering. In the example, the team lead allocates work informally, and nobody checks whether completed analyses were used. There is allocation but no audit.
4. **System 4: who is looking outward and forward?** Ask who tracks changes in the environment: new data sources, shifts in what the departments need, new methods. In many small teams nobody does this deliberately, and the team is surprised by changes it could have seen coming. The analytics group learns about new departmental priorities only when requests arrive.
5. **System 5: what is the team for, and who balances today against tomorrow?** Ask whether the team has a clear identity and whether anyone arbitrates between keeping current work running and investing in change. In the example, the lead is absorbed by System 3 work, and there is no occasion on which the team decides how much capacity to set aside for System 4 concerns.

Having run the five questions, look at the channels between the functions. Does System 3 learn what the units are actually doing, or only what they report? Does what System 4 notices reach the people who set priorities? Then apply the variety test from Chapter 6 in miniature: for each function, ask whether it has enough responses to match the disturbances it faces, and whether the information reaching it has been filtered in ways that hide what matters.

The output of the check is a short list of missing or weak functions and blocked channels. In the composite example, it points to three changes worth trying: someone scheduling the shared engineer's time, a light review of whether analyses were used, and a standing monthly slot for looking ahead. Whether those changes help is an empirical question. Record what you expected to change and look again after a few months, which is more evaluation than most applications of the model have received, and far more than Cybersyn was ever allowed.

Chapter 17. Whole Systems for Public Health: From the Foresight Map to the STOPS Trial

"This analysis by the government's Foresight programme shows that over half of the UK adult population could be obese by 2050." The sentence comes from the executive summary of *Tackling Obesities: Future Choices*, the report of a British government foresight project published in 2007 (Butland et al. 2007). It was widely quoted, frequently misquoted as a prediction, and attached to one of the most reproduced images in the whole systems literature: a sprawling map of the causes of obesity, drawn as a web of interlocking loops.

This chapter follows the idea behind that map through public health from the early 2000s to the present. The idea is that problems such as obesity arise from many interacting causes, that tackling them one cause at a time will fail, and that what is needed is a whole-systems approach. The chapter moves in rough chronological order through five stages: the wave of advocacy that made the case; the Foresight report as a policy event; the systematic reviews that counted what the field had actually produced; a quarrel about what counts as evidence for complex interventions; and the one rigorous trial of a community systems intervention that anyone has run. Along the way it takes a detour into the COVID-19 pandemic, where one family of systems models earned real credit.

The chapter is, in one sense, the book's argument in miniature. The diagnosis at its root is sound, the maps it produced were influential, and the evidence that any of it changed health outcomes was thin for a long time. Then someone ran a proper trial, and the result was neither the triumph the advocates hoped for nor the failure a sceptic might expect.

An idea finds its advocates

The complexity argument reached health care early. In 2001 Plsek and Greenhalgh introduced clinicians to the idea of complex adaptive systems in the *BMJ*, and argued that such systems are better guided by minimum specifications and adaptive approaches than by detailed plans (Plsek and Greenhalgh 2001). The paper was addressed to health care rather than population health, but it set a tone that the public health literature would take up.

The wave broke in 2006. In a single issue of the *American Journal of Public Health*, several papers made the case for systems science in public health. John Sterman argued that public health policies meet resistance because the feedback through which we learn in complex systems is weak, slow and confounded, and he proposed simulation as a complement to conventional evidence-based methods (Sterman 2006). Homer and Hirsch set out the case for system dynamics in public health, emphasising accumulation, feedback and policy resistance, and reviewed applications to date (Homer and Hirsch 2006). Trochim and colleagues reported a concept-mapping study in which 133 public health professionals identified the practical barriers to using systems thinking and modelling (Trochim et al. 2006). The same issue carried a diabetes model built for the US Centers for Disease Control and Prevention, to which this chapter returns below (Jones et al. 2006).

Over the following years the argument spread across the field. Galea and colleagues contrasted single-cause epidemiology with complex-systems approaches and discussed when each is appropriate (Galea, Riddle and Kaplan 2010). Diez Roux identified the features of health-disparities problems, such as feedback, heterogeneity and dynamics across levels, that make complex-systems approaches useful (Diez Roux 2011). Luke and

Stamatakis reviewed three families of systems science method in public health: system dynamics, network analysis and agent-based modelling (Luke and Stamatakis 2012). In global health, a WHO flagship report edited by de Savigny and Adam proposed ten steps to systems thinking for designing and evaluating health-system interventions (de Savigny and Adam 2009), and an editorial by Adam and de Savigny called for a paradigm shift toward systems thinking in strengthening the health systems of low- and middle-income countries (Adam and de Savigny 2012). Paina and Peters used complex adaptive systems ideas to explain why efforts to scale up health services often fall short (Paina and Peters 2012), and Peters asked directly what systems thinking adds to global health and which of its many tools are most useful (Peters 2014).

Obesity became the flagship problem. Lee and colleagues argued that obesity is a systems problem and described the modelling approaches that could address it (Lee et al. 2017). The Lancet Commission report by Swinburn and colleagues framed obesity, undernutrition and climate change as a global syndemic with shared systemic drivers (Swinburn et al. 2019). Knai and colleagues applied systems thinking to the commercial determinants of health, the ways in which industry actors shape the systems that produce health (Knai et al. 2018).

Alongside the advocacy came models, and some of them produced exactly the kind of counterintuitive result that system dynamics had always promised. The CDC diabetes model found that better control of diabetes could raise the prevalence of the disease, because people with well-controlled diabetes live longer, and that long delays separate primary prevention from its effects on outcomes (Jones et al. 2006). A model built for a county health department compared interventions against cardiovascular disease over time (Hirsch et al. 2010). The HealthBound simulation of the US health system projected that combining behavioural and environmental protection with coverage and care would save more lives and cost less over 10 to 25 years than coverage and care alone (Milstein et al. 2011). More recently, a participatory model of youth suicidal behaviour in an Australian region projected that a combination of interventions would have the largest impact over 20 years (Occhipinti et al. 2021).

The diabetes result shows what a stock-flow view can reveal, and a small illustration makes the mechanism plain. The numbers are invented and are not taken from the CDC model. Treat the number of people living with diabetes in a population as a stock. It rises with new diagnoses and falls as people with the disease die. Suppose 10,000 new cases are diagnosed each year and 5 per cent of those living with the disease die each year. The stock settles where outflow equals inflow, at 200,000 people, because 5 per cent of 200,000 is 10,000. Now suppose better management of the disease cuts the death rate among people with diabetes to 4 per cent, with no change in new cases. The stock now settles at 250,000. Prevalence has risen by a quarter, and the rise is a sign of success, because it means people with diabetes are living longer. A health service that judged its diabetes programme by prevalence would conclude that the programme was failing. Only primary prevention, reducing the inflow, brings the stock down, and the model's second finding, that long delays separate prevention from outcomes, explains why that route is so slow to show results.

These are interesting and often sophisticated models, and the diabetes result in particular is a genuine insight of the kind that simple reasoning misses. But every one of them reports simulated outcomes. None compares what happened after a policy informed by the model with what happened where no such policy was adopted. The broader claim that system dynamics modelling improves health policy decisions and outcomes therefore rests on projections, and projections are forecasts of what a model says, not observations of what the world did.

2007: the map that became a policy event

The Foresight project on obesity was run by the UK Government Office for Science. Its report, in the words of its own text, "has involved over 300 experts and stakeholders" (Butland et al. 2007). Its central judgment was that obesity is determined by a complex web of causes "where no single influence dominates" (Butland et al. 2007).

The figures that made headlines were extrapolations. The report's section on trends projected past increases forward and arrived at the possibility that 60 per cent of men and 50 per cent of women aged 21 to 60 could be obese by 2050, and, over a shorter horizon, that some 40 per cent of Britons could be obese by 2025 (Butland et al. 2007). It estimated that the costs to the National Health Service attributable to overweight and obesity were "projected to double to £10 billion per year by 2050", and that the wider costs to society and business could reach £49.9 billion a year at the prices of the day (Butland et al. 2007).

The word that matters in all of these statements is "could". The report extrapolated trends to show where they would lead if nothing changed. It did not forecast that they would continue, and its purpose was precisely to argue for action that would prevent them from continuing. Popular accounts that turned "could be obese" into "will be obese" misrepresented the report, and they also misrepresented its logic, which was the logic of a warning.

The report's own intermediate figures show how the extrapolation worked. Carrying forward the trends of the preceding decades, it arrived at around 36 per cent of men and 28 per cent of women obese by 2015, rising to 47 per cent and 36 per cent by 2025, and on toward the 2050 figures (Butland et al. 2007). An extrapolation of this kind answers a conditional question: if the forces that produced the past trend continue unchanged, where does the trend go? It makes no claim that those forces will continue, and it contains no model of how people, markets or governments might respond as obesity rose. That is exactly the kind of feedback a systems view insists on, which gives the headline figures a certain irony. The most quoted numbers attached to the most famous systems map were produced by a method with no feedback in it at all. That was a reasonable choice for a warning, and a poor basis for anyone who read the numbers as a forecast.

The causal map that accompanied the report, whose construction and limits as a diagram Chapter 10 describes, is a qualitative map rather than a simulation model (Vandenbroeck, Goossens and Clemens 2007). The question here is what the map did as a piece of policy.

Its influence on the way people talk about obesity is hard to overstate. Finegood, Merth and Rutter used it to argue that solutions to childhood obesity need action at several levels of intervention, which they named in terms adapted from Meadows's leverage points (Finegood, Merth and Rutter 2010). The map came to be cited as the iconic picture of what a systems view of a public health problem looks like.

Influence on talk is not the same as influence on policy, and when researchers looked at policy, they found less than the map's fame suggested. Johnston, Matteson and Finegood coded the recommendations in 12 obesity planning documents published between 2004 and 2013 against an intervention-level framework derived from Meadows, and found that most recommendations targeted the lower, less transformative levels (Johnston, Matteson and Finegood 2014). Chapter 12 uses the same study to ask whether Meadows's ranking can be tested; here its significance is simpler. Years after the map, the plans that governments actually wrote still looked much as plans had looked before it. Twelve years after the Foresight report, Public Health England published a guide to help local authorities adopt a whole-systems approach to obesity, laid out as a six-phase process running from setting up, through understanding and mapping the local system, to action, building a network and reflection, and accompanied by a learning report from pilot councils (Public Health England 2019). The guide is useful practical material. It is implementation guidance, not outcome evidence, and its existence implies that the systemic action the map had seemed to call for had not followed from the map itself.

The claim that system maps such as the Foresight map change policy gets a D. The map is widely cited as iconic, but there is no evidence of policy effects that can be attributed to it, and the evidence that exists, from the coding of later plans, points toward business much as usual. That verdict is about attribution. It does not say the map was useless, and it may well have changed how a generation of public health workers think. It says that nobody has shown the map changed what governments did.

What the reviews found

As the literature grew, researchers began to count it, and the counts tell a consistent story.

Carey and colleagues reviewed 117 articles on systems science and systems thinking in public health and sorted them into four categories, running from commentary to formal system dynamics modelling. They found weaknesses in transparency and in how the modelling was carried out (Carey et al. 2015). Chughtai and Blanchet's bibliometric review of 557 publications on systems thinking in public health found near-exponential growth, with half of the publications appearing after 2010, clustered into distinct research communities (Chughtai and Blanchet 2017). Rusoja and colleagues reviewed 516 full texts on systems thinking and complexity in health published between 2002 and 2015, mapped the terms and methods they used, and found no consensus (Rusoja et al. 2018).

When reviewers asked the harder question, whether systems modelling actually works for health policy, the evidence nearly vanished. Atkinson and colleagues' systematic review of system dynamics modelling in support of health policy found six papers describing eight case studies, and no analytic studies of whether such modelling works (Atkinson et al. 2015). Darabi and Hosseinichimeh's review of system dynamics in health and medicine from 1960 to 2018, the first systematic review of its kind, found wide application but considerable variation in the quality of documentation, use of data and validation (Darabi and Hosseinichimeh 2020). And Bagnall and colleagues' systematic review of whole-systems approaches to obesity and other complex public health challenges, covering 65 articles of which 33 concerned obesity, found substantial heterogeneity; positive outcomes were reported, but the designs were weak, and whole-systems thinking was only sometimes evident in interventions that claimed it (Bagnall et al. 2019).

The pattern across these reviews is consistent enough to state firmly. Systems thinking in public health has grown fast and remains mostly conceptual, with much advocacy and comparatively little formal modelling or evaluation. The evidence for that description is bibliometric and consistent across independent teams, and I know of nothing that contradicts

it. The last of Bagnall's findings is the one I would underline. If interventions labelled as whole-systems approaches often do not display whole-systems thinking, then even positive results from them cannot be credited to the systems idea. An evaluation needs to measure not only whether an intervention worked but whether it was systemic in the way it claimed to be.

Why did so much conceptual work produce so little evaluation? The reviews do not answer that question directly, and I offer only a judgment. Part of the answer is visible in Trochim's study, whose respondents identified practical barriers to systems thinking and modelling in their own work (Trochim et al. 2006). Part is structural. A conceptual paper or a system map can be produced by a small team in months; an outcome evaluation of a community-wide programme needs years, many sites and a budget that few funders provide for a method whose value has not yet been shown. The result is a circle in which the absence of evidence makes evaluation hard to fund, and the absence of evaluation keeps the evidence absent. Breaking such a circle takes a deliberate investment by someone willing to risk an unflattering answer, which is part of what makes the trial described below so valuable.

A quarrel about evidence

By the late 2010s the field had reached an uncomfortable position. Advocates were confident that systems approaches were needed, and the reviews showed almost no evidence that they worked. One response was to question the standard of evidence being applied.

In 2017 Rutter and colleagues argued in *The Lancet* for a complex-systems model of evidence in public health, one that goes beyond the linear paradigm built around randomised controlled trials (Rutter et al. 2017). Greenhalgh and Papoutsi argued that health services research still relied on closed-system, predictive designs and needed an overdue shift toward methods suited to complexity (Greenhalgh and Papoutsi 2018). Hawe, Shiell and Riley had earlier proposed treating interventions as events in complex ecological systems, unfolding in settings and networks over time, with

consequences for how they should be evaluated (Hawe, Shiell and Riley 2009). Petticrew and colleagues, in a paper commissioned by WHO, considered what a complexity perspective means for systematic reviews and guidelines, including how questions are framed and what evidence counts (Petticrew et al. 2019).

Other contributors took a more measured line. Moore and colleagues argued that evaluations of whole systems may be neither achievable nor necessary, and that a systems lens can still guide the design of evaluations, for example by identifying the key uncertainties and choosing the timing of follow-up (Moore et al. 2019). McGill and colleagues reviewed 21 qualitative process evaluations that took a complex-systems perspective and proposed a framework for such evaluations (McGill et al. 2020). The National Institute for Health Research's School for Public Health Research issued practical guidance on planning evaluations of local systems approaches in March 2019 (Egan et al. 2019). And in 2021 the Medical Research Council's framework for developing and evaluating complex interventions, updated with the NIHR, explicitly incorporated a complex-systems perspective alongside questions of whether interventions work (Skivington et al. 2021).

This debate is healthy, and much of it is sensible. Nobody should evaluate a community-wide programme as if it were a pill. But the claim that complex interventions require new evaluation methods is a consensus of frameworks rather than a finding, and there is limited evidence that the new methods lead to better decisions. The debate also contains a risk. If the conclusion drawn from the absence of evidence is that the usual standards of evidence do not apply, the systems idea becomes unfalsifiable, and that is the position from which the leverage illusion can never be corrected.

The strongest rebuttal to that risk came from the field itself, in the form of a trial showing that a rigorous design was possible.

Ten communities in Victoria

The WHO STOPS trial, the Whole of Systems Trial of Prevention Strategies for Childhood Obesity, was designed by Allender and colleagues and published as a protocol in 2016 (Allender et al. 2016). Its roots lay in

earlier work in which group model building in a rural Australian community produced a causal loop diagram of the causes of childhood obesity to guide local action (Allender et al. 2015). Chapter 11 describes group model building and those diagrams; the trial used the method to help communities build their own maps and plans.

The design was a stepped-wedge cluster randomised trial in ten communities in Victoria, Australia. The idea of a stepped-wedge design is that every cluster, here every community, eventually receives the intervention, but the order in which clusters start is randomised. At any given time some communities have begun and others have not, so the ones still waiting serve as comparisons for the ones already under way. The design suits interventions that everyone expects to be beneficial, or that cannot be delivered everywhere at once, because nobody is permanently denied the programme. It also allows the effect to be estimated as the interaction between being an intervention community and the passage of time.

A hypothetical schedule shows how this works in practice; it does not reproduce the STOPS timetable. Imagine six towns and three measurement rounds a year apart. Before the first round, the order of starting is drawn at random. Two towns begin the programme straight after the first round, two after the second, and the last two after the third. At the first round, no town has started, so all six provide a baseline. At the second round, two towns have been running the programme for a year and four have not. At the third round, four have started and two are still waiting. Comparing the change in towns that have started with the change in towns that have not, round by round, gives an estimate of the effect that is protected against the most common rival explanation, which is that everything was changing anyway. Because the order was random, there is no reason to expect the early starters to differ systematically from the late ones.

The design has limits, and they are general features of cluster trials rather than criticisms of STOPS in particular. With only ten communities, chance differences between them can be large, and a real but modest effect may be hard to distinguish from noise. Communities that are waiting may not stay untouched, because ideas travel and neighbouring towns share media and

institutions. And a community-wide programme is delivered differently in each place, so the trial estimates the average effect of a family of related interventions rather than of one fixed treatment. None of these limits undermines the trial. They explain why a single well-run trial, however good, leaves questions open.

What the study actually did, as reported in the four-year outcomes paper (Allender et al. 2021), was measure children's behaviours, health-related quality of life and body mass index in the participating communities over time, with roughly 1,800 to 2,400 children measured in each wave. The weight outcome was the BMI z-score, a measure of body mass index adjusted for a child's age and sex.

The results were mixed in an informative way. The interaction between intervention and time for BMI z-score was statistically significant. But the pattern behind that interaction was not a steady improvement. In the intervention communities, BMI fell between 2015 and 2017 and then rose again between 2017 and 2019. The authors, to their credit, did not overclaim. Their conclusions claimed benefits for takeaway food intake and for quality of life, not for weight (Allender et al. 2021).

This is the most rigorous test the field has run, and it deserves full credit on several counts. It published its protocol. It randomised. It measured objective outcomes in thousands of children. It reported a result that did not fit a simple success story, and its authors confined their claims to what the data supported. Anyone who argues that systems interventions are too complex to evaluate properly has to reckon with the fact that this one was evaluated properly.

What does the trial tell us? The claim that whole-systems or community systems interventions reduce childhood obesity gets a C. There is one rigorous trial, and its weight results are mixed; the supporting literature, as Bagnall's review showed, is low in quality. The trial shows that community systems interventions can change some behaviours and wellbeing outcomes. Whether they reliably reduce obesity remains open, and so does the question of which components of a systems intervention do the work. A descriptive study of a successful early-childhood obesity initiative, built

from 16 stakeholder interviews, identified six feedback loops involving collaboration, networks and awareness (Owen et al. 2018). Such accounts are useful for generating hypotheses about mechanisms. They cannot test them.

The fall and rise in BMI invites speculation, and I will resist most of it, because the trial was not designed to explain it. It is worth noting only that a pattern of early improvement followed by reversal is a familiar shape in systems work, and that it is exactly the shape a short evaluation would miss. A trial that stopped after two years would have reported a clear weight benefit. The four-year follow-up is what revealed the reversal, and that is an argument for long follow-up in any evaluation of this kind.

Models that earned their keep in the pandemic

The COVID-19 pandemic gave systems modelling a test it had rarely faced: forecasts made in public, compared with outcomes that arrived within weeks. Currie and colleagues set out early in the pandemic how operational simulation methods, including discrete-event, system dynamics and agent-based models, could support decisions beyond what epidemiological models offered (Currie et al. 2020).

The most striking results came from models that included feedback from behaviour. Rahmandad, Lim and Sterman built a behavioural system dynamics model of COVID-19 and estimated it for 92 nations (Rahmandad, Lim and Sterman 2021). The model treated people's responses to perceived risk as part of the system, so that protective behaviour could rise and fall with the perceived danger. The estimates suggested that cumulative cases and deaths were 7.03 and 1.44 times the official reports respectively, and that without adherence fatigue, the gradual erosion of protective behaviour over time, cases would have been 47 per cent lower. Differences in how responsive populations were to risk explained large differences in deaths between countries.

A second study asked the forecasting question directly. Rahmandad, Xu and Ghaffarzadegan examined the COVID-19 forecast models submitted to the CDC and found that better long-term accuracy went with three modelling choices: representing transmission mechanisms, including endogenous

behavioural responses, and resetting state variables (Rahmandad, Xu and Ghaffarzadegan 2022). A simple model with these features, of a type the authors called SEIRb, matched the best models in the collection up to 20 weeks ahead.

This is a rare case of a system dynamics claim tested quantitatively against real forecasts, and it deserves credit. The claim that models including endogenous behavioural feedback forecast better gets a C, moving toward B. The evidence comes from one epidemic and one team's analysis, and it needs replication in other domains before it can be treated as general. But the direction is clear, and the finding has an attractive feature: the advantage came from the right mechanisms, not from size or complexity. A small model with behaviour in it did as well as the best large ones. Chapter 20 returns to this point in its discussion of small, transparent models.

It is not hard to see why behaviour matters so much for an epidemic forecast. A model without behavioural feedback projects that infections will keep growing at the current rate until immunity builds up, so it tends to overshoot during a surge and to miss the turning point. A model in which rising danger prompts people to reduce their contacts contains a balancing loop that bends the curve before immunity does, and a model with adherence fatigue contains a slow erosion of that loop, which helps explain why waves recur. The mechanisms are the same ones Part I of this book showed people misjudge: accumulation, delay and feedback. What the COVID models demonstrated is that putting those mechanisms into a forecasting model, and estimating them against data, improves the forecast. That is a much more modest claim than the grand promises of systems modelling, and a much better evidenced one.

The caution that applies is the one Saltelli and colleagues pressed in their manifesto for responsible modelling during the same period: attend to the assumptions, the framing and the unknowns, and resist the hubris that a good fit invites (Saltelli et al. 2020). A model that forecast well in one epidemic has shown something important about that epidemic. Whether

behaviour-rich models forecast better in other domains, such as consumer demand or health service use, is a question the COVID evidence raises and does not answer.

Building the evaluation in before the intervention

The story of whole systems in public health ends with a practical lesson. The field spent more than a decade producing maps, models and frameworks with almost no outcome evidence, and when a proper trial was finally run it produced the most useful knowledge the field has. The obvious conclusion is to design the evaluation at the same time as the intervention, not afterwards. The plan below is built on the MRC framework for complex interventions (Skivington et al. 2021) and the NIHR guidance on evaluating local systems approaches (Egan et al. 2019), and on the design of the STOPS trial. It is written for a local systems intervention of moderate scale, such as a council's healthy-weight programme.

1. **Write down the outcomes before you begin.** Separate the proximal outcomes that the process should change quickly, such as shared understanding among partners, new collaborations or changes in local policy, from the distal outcomes that matter most, such as behaviours, health measures or service use. Specify how each will be measured and when. The STOPS authors' decision to report weight honestly, when it did not move as hoped, was possible only because weight was named in advance.
2. **Choose a comparison design.** If the intervention will be rolled out to several areas anyway, randomise the order and use a stepped-wedge design, which gives every area the programme and still provides a comparison. If randomisation is impossible, identify comparison areas that are not receiving the intervention and measure them in the same way at the same times. A before-and-after study in the intervention area alone will not distinguish the programme's effect from everything else that happened in the same years.

3. **Plan for long follow-up.** Systems interventions work through feedback and delay, and their effects may appear, fade or reverse over years. The STOPS reversal between years two and four is the example to keep in mind. Budget for at least one measurement well after the expected effect.
4. **Measure how systemic the intervention actually was.** Bagnall's review found that interventions called whole-systems approaches often did not display whole-systems thinking. Decide in advance what would count as systemic practice in your setting: for example, the number of sectors actively involved, whether actions address several levels of the problem at once, and whether the plan changes in response to what is learned. Track these as process measures, using qualitative process evaluation of the kind McGill and colleagues describe where numbers will not capture what is happening.
5. **Record the map as a hypothesis.** If the intervention begins with a system map or a group model, write down which loops the partners expect their actions to affect, and what change in which measure would show that they did. Revisit the map at each measurement point. A map that is never checked against what happened is a picture, and the practice of checking it is what turns it into evidence.
6. **Say in advance what would count as failure.** State what result on the distal outcomes would lead you to conclude that the intervention did not work as intended, and commit to publishing it. This is the step most often skipped and the one that protects an evaluation from being reinterpreted after the fact.

Consider how the plan might look for a hypothetical council; the example is a composite and describes no real authority. The council intends to run a healthy-weight programme in eight neighbourhoods over three years, starting with group model building in each. Under the plan, it names two proximal outcomes (partners' agreement on priorities, measured by a short survey before and after the workshops, and the number of joint actions started across sectors) and two distal ones (the proportion of primary school children above a healthy weight, taken from routine measurement, and children's reported daily consumption of sugary drinks). It randomises the

order in which neighbourhoods start, in three groups. It commits to a final measurement in year five, two years after the last group starts. It records, for each neighbourhood, how many sectors are active, whether actions span schools, food outlets and transport, and whether the local plan is revised after each measurement round. It lists the three loops on each neighbourhood's map that the partners expect to shift. And it states that if the distal measures in started neighbourhoods do not differ from those in waiting neighbourhoods by year five, it will report the programme as ineffective on those measures. Whatever the result, the council will know more at the end than the field knew about whole-systems approaches for most of the period this chapter describes.

None of this requires an unusual budget or exotic methods. It requires deciding, before the first workshop, that the question of whether the intervention worked will be answered by evidence collected for the purpose. The Foresight map shows how far a compelling picture can travel without that decision. The STOPS trial shows what becomes possible once it is made.

Chapter 18. The Commons Are Not Tragic: Ostrom's Evidence and the Misperception Twist

What happens to a pasture, a fishery or an aquifer that many people use and nobody owns? For a long time the standard answer in economics, ecology and policy was that it would be ruined. Each user gains the full benefit of taking a little more and bears only a fraction of the cost, so each takes more, and the resource collapses. The answer seemed to follow from the structure of the situation as surely as overshoot follows from growth and delay, and it came with a lever attached: take the resource out of common hands, either by giving it to private owners or by placing it under the control of the state.

This chapter is built on two contrasts. The first sets that standard answer against the evidence that Elinor Ostrom and her colleagues assembled from real communities over several decades. The second sets the field against the laboratory, and it contains a twist. Ostrom's field evidence showed that communities can often govern shared resources well, and Moxnes's laboratory evidence showed that even a single owner with no one to compete against can ruin a resource by misreading its dynamics. Taken together, the two contrasts say something precise about where leverage over shared resources lies, and it is quite different from what the standard answer supposed.

This is the chapter in which the book's credit is most complete. Ostrom's work is the clearest example in these pages of systems-level claims about governance tested properly: developed from careful observation, stated so that they could be wrong, and then checked by other researchers across many cases.

Hardin's parable and Ostrom's evidence

The standard answer had a famous source. In 1968 Hardin published "The Tragedy of the Commons" in *Science* (Hardin 1968). His argument was a parable rather than a study, built around users of a resource open to all, each of whom gains the full benefit of taking a little more while sharing the cost of the damage with everyone else. Resources open to all, Hardin concluded, are overexploited, and the remedy is coercion or privatisation.

The parable has great force as a piece of reasoning about incentives. Its weakness was the leap from a possible outcome to an inevitable one. The users in the parable do not talk to each other, make rules, watch each other or punish those who cheat. They act alone, each responding only to private gain. Real users of shared resources often do all of those things, and whether they do them well is an empirical question that the parable answers by assumption.

The arithmetic of the parable is easy to reproduce with invented numbers. Suppose ten users share a resource, and one of them considers taking an extra unit that is worth 10 to whoever takes it. Suppose also that the extra take damages the resource by 15 in total, a loss shared equally among all ten users. The user who takes the extra unit gains 10 and bears only 1.5 of the damage, so taking it looks worthwhile. But if all ten reason the same way, each gains 10 and bears 15, and everyone ends up worse off. That is the dilemma in its simplest form, and nothing about it is wrong as reasoning. What the arithmetic leaves out is everything the users might do about it: agree to limit their take, watch one another, impose a penalty on anyone who breaks the agreement, and revise the agreement when conditions change. Ostrom's research was a long investigation of whether and how real users do those things.

Ostrom set out to answer it with evidence. *Governing the Commons*, published in 1990, was a comparative institutional analysis of long-lived regimes for managing common-pool resources, including irrigation systems, fisheries and communal lands (Ostrom 1990). A common-pool resource, in the sense Ostrom used, is one from which it is costly to exclude potential users and in which one user's take reduces what is left for others. Many such

regimes had lasted for very long periods under the management of their users, without either privatisation or central control. Ostrom asked what distinguished the regimes that endured from those that failed, and her answer took the form of eight design principles that characterised the robust ones.

Several features of Ostrom's approach deserve emphasis, because they are what made the work so much stronger than most of what this book has examined. She began from cases rather than from a model, and she compared many of them instead of generalising from one. She stated her conclusions as principles that could be checked against further cases, and she treated them as characteristics of robust institutions, not as a recipe guaranteed to produce them. In later work she built a framework and a typology of rules for analysing institutions, drawing on both field and laboratory studies (Ostrom 2005), and proposed a general, multi-tier framework of variables for diagnosing social-ecological systems (Ostrom 2009).

She was also emphatic about what her findings did not license. With Janssen and Anderies she argued against panaceas, the habit of applying a single blueprint solution to every environmental problem regardless of its particulars (Ostrom, Janssen and Anderies 2007). In her Nobel lecture she set out the case for polycentric governance, with many centres of decision operating at different scales, and marshalled laboratory and field evidence against the idea that the only choice is between markets and states (Ostrom 2010).

With Dietz and Stern she also warned that governing the commons is a struggle and not a solved problem (Dietz, Ostrom and Stern 2003). Locally evolved institutions can sustain commons for centuries, their review found, but they often fail under rapid change, and the requirements for governance that adapts to change are demanding. This is a finding in both directions. It refutes the claim that commons are doomed, and it refutes the romantic counter-claim that communities left alone will always manage well.

The claim that the tragedy of the commons is inevitable without privatisation or state control gets a D, and the D here means contradicted rather than merely untested. Extensive field evidence from Ostrom's comparative work, laboratory experiments, the Dietz review and the later meta-review of the design principles show that self-governance often works. Hardin's parable remains a useful description of one possible outcome, the one that follows when users cannot communicate, make rules or enforce them. As a claim about what must happen, it does not survive.

Covenants in the laboratory

Field comparisons have a known weakness. Communities that manage their resources well may differ from those that fail in many ways besides their institutions, and a comparison of cases cannot fully separate the institutions from everything else. Ostrom's response was to take the question into the laboratory as well.

In "Covenants with and without a Sword", Ostrom, Walker and Gardner reported experiments in which groups of participants harvested from a shared resource under controlled conditions (Ostrom, Walker and Gardner 1992). Allowing participants face-to-face communication, with or without a sanctioning mechanism, substantially improved outcomes. The title captures the finding: covenants, agreements among the users themselves, could sustain cooperation without the sword of external enforcement. Self-governance, the authors concluded, is possible.

Later work built richer laboratory environments. Janssen, Holahan, Lee and Ostrom introduced a laboratory environment in which a shared resource changes over both space and time, and reported experiments on communication and sanctioning within it (Janssen et al. 2010). The aim was to bring the laboratory closer to the dynamics of real social-ecological systems while keeping the control that allows causes to be identified.

The field and the laboratory, set against each other, strengthen one another. The field shows that the principles describe real institutions that lasted. The laboratory shows that the mechanisms the principles point to, communica-

tion and mutual enforcement among users, cause better outcomes when other differences are held constant. Neither kind of evidence alone would carry the conclusion as far as the two together.

Eight principles under test

The strongest test of any set of principles derived from cases is whether they hold up in cases their author did not choose. Two reviews have put Ostrom's principles to that test.

Cox, Arnold and Villamayor Tomás reviewed 91 studies that had evaluated the design principles, explicitly or implicitly, and concluded that the principles are well supported empirically (Cox, Arnold and Villamayor Tomás 2010). Their review, read in full for this chapter, is worth describing in some detail, because it shows what well supported means in practice. The principles, as they coded them, concern:

1. well-defined boundaries, both around the community of users and around the resource itself;
2. congruence between the rules for taking from the resource and providing for it, and between those rules and local conditions;
3. collective-choice arrangements, so that most people affected by the operating rules can take part in changing them;
4. monitoring, which they split into the presence of monitors and the condition that monitors are members of the community or accountable to it;
5. graduated sanctions, which rise with the severity or repetition of a violation;
6. low-cost mechanisms for resolving conflict;
7. minimal recognition by external authorities of the users' right to organise their own institutions;
8. nested enterprises, with governance activities organised in several layers.

The support was not uniform. The reviewers found strong evidence for boundaries around the user community and more moderate evidence for boundaries around the resource; very strong evidence for congruence with local conditions; moderate evidence for the presence of monitors but very strong evidence that monitors should be accountable to the users; and moderate support for collective choice, graduated sanctions, conflict resolution, recognition of rights and nesting. They also found that the most serious criticisms of the principles in the literature were abstract rather than empirical, and that the most important empirical criticism was that the principles are incomplete: group size, differences among users and the wider political regime also matter in many cases. They recommended a probabilistic rather than deterministic reading of the principles, and said they remained uncertain whether the principles apply at every scale (Cox, Arnold and Villamayor Tomás 2010).

Baggio and colleagues then used qualitative comparative analysis on 69 cases of irrigation, fishery and forestry management to ask which combinations of principles are associated with success and failure (Baggio et al. 2016). They found that which principles matter depends on the kind of resource and its infrastructure, but that two matter everywhere: congruence with local conditions, and a proportional relationship between the costs users bear and the benefits they receive. That finding sharpens the principles considerably. It suggests that the eight are best read as a set of conditions whose importance varies with context, with a core of fit and fairness that is needed in every setting.

Should the principles be generalised beyond shared natural resources? Wilson, Ostrom and Cox argued that the core design principles apply to groups in general, including organisations, communities and teams (Wilson, Ostrom and Cox 2013). The argument is interesting, and it may prove right. At present the extension is theoretical, and evidence outside common-pool resources is limited.

A composite example shows how the principles work together; it is invented for illustration and draws on no particular case. Picture a farming community that shares an irrigation canal. The community knows who its

members are and which fields the canal serves, which answers the question of boundaries. Water is allocated by turns, and the turns are adjusted in dry years, so the rules fit local conditions; members contribute labour to maintaining the canal in proportion to the land they irrigate, so costs follow benefits. Rule changes are agreed at a meeting of members. Because turns are taken in sequence along the canal, neighbours can see whether anyone takes water out of turn, so monitoring costs almost nothing and is done by the users themselves. A first offence brings a small fine, repeated offences larger ones. Disputes go to an elected member who settles them quickly. The regional authority accepts the community's right to set its own rules, and the community's canal is one branch of a larger system with its own coordinating body. No single one of these features explains why such a regime might last. The principles describe a configuration, and Baggio's analysis suggests that a missing piece matters more in some configurations than in others.

The grade for the design principles therefore splits. For common-pool resource institutions I would grade the claim that the principles characterise robust regimes B: it is supported across many cases, by independent reviewers, with a laboratory programme behind the mechanisms, but the supporting studies are largely correlational comparisons, and the QCA shows that the principles' importance depends on context. For the generalisation to all groups the grade is D, because it has not been tested. Neither grade diminishes the achievement. A B for a governance claim derived from field cases is rare in the literature this book covers, and it was earned by exactly the kind of work the book argues for.

For the leverage illusion, the lesson is specific. Ostrom did not find a single lever for governing the commons. She found a set of institutional conditions, discovered in the field, tested across many cases and shown to vary in importance with context. That is what grounded leverage looks like. It is less dramatic than a single structural intervention and far more reliable, because each element has been checked against what actually happened in real communities.

It is worth setting this beside the other cases in this part of the book. World3 identified a lever from the structure of a model, and history never tested it. Cybersyn built a lever from a theory of regulation and was stopped before it could be judged. The Foresight map pictured a web of causes and left no trace that anyone could attribute in the policies that followed. Ostrom's principles came from the other direction: from institutions that had already worked, compared with others, and then tested by researchers who had no stake in her being right. Chapter 12 anticipated this contrast when it listed well-evidenced institutional rules among the few places where the language of leverage can be used with a straight face. The difference lies in the order of operations. Ostrom looked at outcomes first and inferred structure from them. The other cases reasoned from structure toward outcomes, and none of their claims was ever tested in the way her principles were.

The owner who overfishes

Now the twist. Hardin's remedy for the commons was to remove the commons, by privatisation or state control, so that a single decision-maker bears the full cost of overuse. If the problem is only one of incentives, that should work. A sole owner has every reason to manage the resource for the long term.

Chapter 5 described the experiments in which Moxnes tested that reasoning, and they need only a brief recall here. Participants from the Norwegian fisheries sector, each given sole ownership of a simulated fish stock, still built too much fishing capacity and ran their stocks down (Moxnes 1998), and later experiments designed to rule out the commons problem found the same pattern, which Moxnes attributed to misperception of the stocks, flows and nonlinearities involved (Moxnes 2000). The resource was overused without any commons dilemma at all.

A hypothetical with invented numbers shows how a sole owner can go wrong without any competitor. Suppose a fish stock can sustain a harvest of 100 tonnes a year indefinitely, and the owner's boats currently land 80. Catches are good and the stock looks healthy, so the owner orders more

boats, which take two years to build. By the time they arrive, the owner's landings have risen toward the sustainable level, and the new boats push the harvest to 130 tonnes. The stock begins to fall, but slowly at first, and catches stay high for a while because the boats are working a stock that is still large. When catches finally drop, the owner has capacity for 130 tonnes and a stock that can no longer sustain even 100. Every decision along the way looked reasonable given what the owner could see at the time. The trouble lay in the delays between ordering capacity and using it, and between overharvesting and seeing its effect, which is exactly the kind of structure Part I showed people misjudge.

Kampmann and Sterman's market experiments, described in Chapter 3, point the same way: market institutions improved performance but left it well below optimal, moderating the misperception of feedback without removing it (Kampmann and Sterman 2014).

The claim that resource overuse can arise from misperception even without a commons dilemma gets a B. The experimental programme is striking and consistent, and nothing in the literature contradicts it. The grade stops short of A because the results come largely from one research group, with modest samples, and replication outside that programme is limited.

The two contrasts of this chapter now meet. Ostrom showed that the incentive problem Hardin described can be solved by institutions that users build for themselves, and that privatisation and state control are not the only remedies. Moxnes showed that solving the incentive problem, by any means, leaves a second problem in place. A sole owner facing no dilemma at all can still overinvest and deplete, because the dynamics of a renewable resource, with its delays between investment and harvest and between harvest and regeneration, are hard to read. The institutions that solve the dilemma of the commons do not, by themselves, solve the problem of seeing the stock.

This is the clearest point in the book at which the two halves of the leverage illusion meet. Part I showed that people misread accumulation and delay. Hardin's lever, privatise, assumed that once incentives were aligned the resource would be managed well, which is to assume that the owner can see

the system. Moxnes's owners could not. Ostrom's principles, meanwhile, include features that may help with the perception problem as well as the incentive problem. Monitoring makes users' behaviour visible to the community, and where it extends to the resource itself it can make the stock visible too. Rules congruent with local conditions embed accumulated local knowledge of how the resource behaves. Whether that is why institutions built on the principles endure is a hypothesis, and as far as I know it has not been tested directly.

Stability against resilience

The last pair of this chapter sets two ideas about what it means for a system to be healthy. In a 1973 paper that founded resilience thinking, C. S. Holling distinguished stability, the tendency of a system to return to an equilibrium after a disturbance, from resilience, the capacity of a system to persist and absorb change while keeping the relationships among its parts intact (Holling 1973). A system can be stable without being resilient: it may return quickly to its equilibrium after small shocks yet be tipped into a different state entirely by a large one. And a system can be resilient without being stable: it may fluctuate widely and still keep its essential structure.

The distinction matters for management. Holling and Meffe later described what they called the pathology of natural resource management (Holling and Meffe 1996). Command-and-control management, they argued, reduces the natural variability of a resource in the pursuit of steady yields. In doing so it erodes the resilience that the variability maintained, and the system becomes brittle and more vulnerable to collapse. The argument is a close cousin of the capability traps and fixes that fail that other chapters describe, and of Scott's critique of high-modernist planning in Chapter 16.

Resilience thinking grew into a substantial field. Walker, Holling, Carpenter and Kinzig defined resilience, adaptability and transformability for social-ecological systems using the image of basins of attraction, in which a system can rest in one of several states; in that picture, resilience reflects how hard it is to push the system out of its current basin (Walker et al. 2004). Folke traced how the resilience perspective emerged and how it came to be used

for the analysis of social-ecological systems (Folke 2006). Levin framed ecosystems and the biosphere as complex adaptive systems (Levin 1998). Liu and colleagues synthesised six case studies of coupled human and natural systems and found nonlinear dynamics with thresholds, reciprocal feedback, time lags, legacy effects and surprise (Liu et al. 2007). Scheffer and colleagues assembled evidence from lakes, coral reefs, oceans, forests and arid lands for alternative stable states and sudden shifts between them, and concluded that management should aim to maintain resilience (Scheffer et al. 2001).

The field's success brought a problem that this book has met before. Brand and Jax documented the proliferation of definitions of resilience and distinguished the descriptive ecological concept from a vaguer, normative use of the word as a boundary object that different communities can rally around without agreeing what it means (Brand and Jax 2007). The ecological concept is precise enough to be tested. The boundary object is not. Chapter 19 meets yet another meaning in safety science, where Woods distinguished four senses of resilience, and it is worth keeping the ecological sense described here separate from those.

Resilience thinking and Ostrom's institutional work converge on a point about governance. If social-ecological systems contain thresholds, lags and surprises, as Liu and colleagues found, then no fixed rule set will suit them for long, and governance must be able to change its own rules as conditions change. That is the requirement for adaptive governance that Dietz, Ostrom and Stern set out, and it explains their finding that locally evolved institutions, robust for centuries under slow change, often fail when change is rapid. It is also the reason Ostrom resisted panaceas. A blueprint that works under one set of conditions may fail under another, and the capacity to notice the difference and adjust is itself one of the conditions that matter. The design principle of collective-choice arrangements, which lets users change their own rules, reads in this light as a mechanism for resilience as well as for fairness.

Warnings that arrive too late or too often

If systems can shift suddenly between states, the obvious practical question is whether the shift can be seen coming. In 2009 Scheffer and colleagues proposed a set of generic early-warning signals for critical transitions (Scheffer et al. 2009). As a system approaches a threshold, the theory goes, it recovers more slowly from small disturbances, a phenomenon called critical slowing down. The slower recovery shows up in data as rising variance and rising autocorrelation, meaning that each observation looks more like the one before it. In principle, a manager watching those statistics could see a transition approaching before it arrived. The idea connects to the planetary tipping points discussed in Chapter 15, where the possibility that climate tipping points are closer than thought has been pressed as a reason for caution (Lenton et al. 2019).

The idea is elegant and has support in some laboratory and field systems. Its practical value is another matter. Boettiger and Hastings showed that common early-warning indicators can have severe rates of false alarms and missed detections, even under favourable assumptions, and they called for routine quantification of the uncertainty in such signals (Boettiger and Hastings 2012). An alarm that sounds often when nothing is wrong, or stays silent when a transition is coming, can do more harm than no alarm at all, because it breeds either complacency or exhaustion.

A hypothetical makes the cost concrete. Imagine an agency monitoring 50 lakes for signs of an approaching shift to a degraded state, and suppose that in a given decade two of them are truly approaching such a shift. An indicator that flags a quarter of lakes that are not approaching a transition would raise about 12 false alarms among the 48 stable lakes, and if it also missed one of the two genuine cases, the agency would face 13 warnings of which one was real. The numbers are invented, but the arithmetic is general: when true transitions are rare, even a modest false-alarm rate swamps the real signals. An agency that responded to every warning would exhaust its budget; one that learned to ignore them would miss the case that mattered. That is why Boettiger and Hastings's call for routine uncertainty quantification is more than a technical nicety.

The claim that early-warning signals can anticipate critical transitions gets a C. It is supported in some systems, and the theory is sound, but the signals are unreliable as general forecasting tools. This matters for the leverage illusion in a particular way. A reliable early-warning signal would be a lever of the most valuable kind, telling managers when to act before a threshold is crossed. The evidence says that such signals exist in principle and cannot yet be relied on in practice.

Auditing a commons, and your own perception of it

The chapter's two contrasts produce two tools, one for the institutions that govern a shared resource and one for the people who manage it. The first rests on B-grade evidence for common-pool resources; applying it to other kinds of group, such as a department sharing a budget, is an extrapolation that has not been tested.

A governance audit using the eight design principles. For a shared resource you are involved in managing, such as a fishery, a groundwater basin, a shared equipment pool or a common budget, work through the eight principles as questions. Score each as present, partly present or absent, and note the evidence for your score.

1. Are the boundaries of the user group and of the resource clear? Could you list who is entitled to use it and what exactly they are entitled to use?
2. Do the rules fit local conditions, and are the obligations users bear roughly proportional to the benefits they receive? Baggio's analysis found these two matter in every setting, so weight them most heavily.
3. Can the people affected by the rules take part in changing them, in practice and not only on paper? Cox's reviewers noted that collective-choice arrangements can exist in form while being captured by powerful local or outside actors.
4. Is someone monitoring both the resource and the users' behaviour, and are the monitors accountable to the users?
5. Do sanctions start small and rise with the seriousness or repetition of a breach?

6. Is there a quick, cheap way to resolve disputes?
7. Do outside authorities accept the users' right to make their own rules?
8. If the resource is part of a larger system, are there layers of governance that connect local rules to wider ones?

Treat the result probabilistically, as the Cox review recommends. Absent principles do not guarantee failure, and a full set does not guarantee success. The principles are also incomplete, so add questions about group size, differences among users and the wider political setting. The audit's value lies in pointing to the gaps most likely to matter.

To see the audit at work, take a hypothetical group of small businesses that share a groundwater aquifer; the group is a composite invented for illustration. Working through the questions, the members find that boundaries are clear, because every licensed well is registered, and that rules are partly congruent, because extraction limits are set per well regardless of how much land each business irrigates, which breaks the proportionality that Baggio found matters everywhere. Collective choice exists on paper through an annual meeting, but the meeting is dominated by the three largest users. Monitoring is weak: meters exist, but readings are self-reported and nobody checks them. There are no graduated sanctions, only the threat of licence withdrawal, which is never used. Disputes go to a regional tribunal that takes months. The regional authority recognises the group's right to set rules, and the group sits within a catchment-wide body. The audit points to four gaps, of which two (proportionality and monitoring) are the ones the evidence most strongly suggests will matter. It does not say the aquifer will be depleted. It says where the group's arrangements differ from those of regimes that have endured.

A misperception check. The audit addresses the dilemma of incentives. It does not address Moxnes's finding that a manager with no dilemma at all can misread the resource. For that, answer four questions before any decision about capacity or harvest.

1. What is the stock, and how do you know its current level? Name the measurement and its date. If the number is an estimate, say how uncertain it is.

2. What are the delays? Estimate how long it takes for an increase in capacity to become an increase in harvest, and how long it takes for the stock to respond to a change in harvest. Moxnes traced his participants' overinvestment to misperceived stocks, flows and nonlinearities, and delays are where such misperceptions hide.
3. What is already committed? Count the capacity on order and the harvest that existing capacity will take if it is used, before adding more, in the same way that Chapter 3's supply-line accounting counts goods already on order.
4. What would you expect to see if you were overharvesting, and when? Write down the signal and the date by which it should appear, and check it when the date arrives.

The two tools together reflect the chapter's conclusion. Governing a shared resource well requires institutions that solve the problem of competing incentives, and Ostrom showed in detail how communities build them. It also requires managers who can see the stock through its delays, and the laboratory shows how often even well-motivated owners cannot. Neither tool is a lever in the illusory sense. Both are ways of checking conditions that the evidence says matter.

Part V. Systems Thinking Without the Illusion

The final part is brief and deliberately practical. It asks what systems thinking looks like once it stops promising levers and settles for being a disciplined way of seeing and a set of claims to be tested. The first of its two chapters follows a field that has gone some way toward that; the second draws the book's argument into a code for practice.

Chapter 19. After the Accident: Safety Science and the Systems Turn

In April 2008, six students and their teacher drowned in the Mangatepopo gorge in New Zealand during an outdoor education activity. Tragedies of this kind are investigated because people want to know why they happened and how to prevent the next one, and the answers depend heavily on how the investigation is done. Later, separate analysts took the same tragedy and analysed it, each using a different method: AcciMap, HFACS and STAMP. When their outputs were compared, the researchers reported "significant differences across the three methods" (Salmon et al. 2012). The same deaths, examined with the same care by different methods, produced different accounts of why they had happened.

That finding could be read as an embarrassment for safety science. This chapter reads it differently. It is evidence of something unusual and admirable: a field that took up systems ideas, built new methods around them, set those methods side by side on the same events, and reported what it found even when the findings were awkward for its newest tools. Safety science is the best example in this book of what an honest use of systems thinking looks like. It changed what investigators look for, it examined its own tools, and it has been candid about what it does not yet know. It has not shown that its systemic methods reduce accidents, and the chapter is equally clear about that. The two conclusions belong together, and holding them together is the point.

From chains of events to latent conditions

The systemic writers defined their approach against an older picture of accidents, one they describe as tracing a chain of events backward from the harm to a failed component or a mistaken action by an operator (Leveson 2004; Dekker et al. 2011). The chain would end at a root cause, usually a person at the sharp end, and the remedy followed: retrain, discipline or replace the person, or fix the part. The systems turn in safety science was, above all, a rejection of that picture as an adequate account of accidents in complex sociotechnical systems.

Perrow's *Normal Accidents*, published in 1984, made an early and influential case (Perrow 1984). Using the Three Mile Island nuclear accident as his anchor case, Perrow argued that some accidents are "normal" in a specific sense: in systems that combine interactive complexity with tight coupling, unexpected interactions between failures are inevitable, and their consequences spread too quickly for operators to understand and contain them. In Perrow's argument, blaming the operators at Three Mile Island misread the event. The accident was a property of the system's design, and similar accidents should be expected in any system with the same combination of features.

Rasmussen extended the systems view upward and outward. His 1997 paper treated risk management as a control problem spanning all levels of society, from the people doing the work through their managers and companies to regulators and government (Rasmussen 1997). He argued that systems migrate toward the boundary of acceptable performance under everyday pressures, so that safety margins can erode gradually through a series of locally reasonable decisions until an accident occurs. That picture became the basis of AcciMap, a method that lays out the contributing factors to an accident across those levels of the sociotechnical system, so that decisions far from the scene can be seen alongside the actions at the sharp end.

Reason's work supplied much of the vocabulary in which the systems view reached a wider professional audience. His *Human Error* integrated cognitive research on types of error, distinguishing slips, lapses and mistakes, and brought that research together with the idea of latent conditions, which in

Reason's account are weaknesses in a system created by decisions made well before an accident, often by people remote from it (Reason 1990). *Managing the Risks of Organizational Accidents* developed the organisational model and the idea of safety culture (Reason 1997). His short 2000 article in the *BMJ* contrasted the person approach to error, which focuses on the individual who made the mistake, with the system approach, which focuses on the conditions under which people work, and it popularised the Swiss cheese model (Reason 2000). In that model, an organisation's defences are pictured as layers, each with weaknesses, and an accident occurs when the weaknesses in several layers happen to line up.

Vaughan's study of the Challenger launch decision gave the systems view its most compelling case study (Vaughan 1996). Reconstructing the decision from archives and interviews, Vaughan introduced the normalisation of deviance: the process by which an organisation comes to accept, step by step, departures from its own safety standards, because each departure is followed by no disaster and so becomes the new baseline. Her account located the causes of the disaster in organisational culture and structure, rather than in individual wrongdoing, and it remains a model of what a systemic case study can achieve.

These four bodies of work share a conviction that accidents in complex systems cannot be understood by tracing a single chain back to a single cause. They differ in emphasis. Perrow stresses system design, Rasmussen migration under pressure, Reason latent conditions and defences, and Vaughan culture. Together they changed what investigators looked for, moving attention from the sharp end, where the operator's action is visible, toward the blunt end, where decisions about design, staffing, training, regulation and production pressure shape what the operator can do.

It is worth asking why the chain-of-events picture was so durable, because the same forces operate wherever systems thinking meets practice. A chain is easy to draw, easy to explain to a board or a court, and it ends at a point where something can be done: a person can be retrained or a part replaced. It also satisfies a natural wish for accountability. The systemic picture is harder on all three counts. It spreads causes across many levels and many

years, it rarely ends at a single point where a fix can be applied, and it can seem to excuse the individuals involved. The systems turn in safety science had to win against those advantages, and the fact that it largely did so in the research literature, if less completely in practice, is part of what makes the field's history instructive.

Control, variability and the newer models

A second generation of systemic methods went further, and it is these that the comparative studies later tested. Leveson's STAMP (System-Theoretic Accident Model and Processes), introduced in 2004, reframed accidents as the result of inadequate control of safety constraints in a sociotechnical control structure, where earlier models had seen chains of component failures (Leveson 2004). The family resemblance to the cybernetics of Chapter 6 is plain: the organisation is modelled as a control structure whose controllers must enforce safety constraints, and an accident is a failure of that control. *Engineering a Safer World* developed the approach at book length, adding STPA (System-Theoretic Process Analysis), a hazard analysis method for use in design, and CAST (Causal Analysis based on Systems Theory), a method for analysing accidents after they occur (Leveson 2011). Leveson later proposed leading safety indicators built on STAMP, specific to each system and based on the assumptions its safety depends on (Leveson 2015).

Hollnagel took a different route. His Functional Resonance Analysis Method, FRAM, models everyday performance variability in sociotechnical systems and how that variability can combine, or resonate, to produce unexpected outcomes (Hollnagel 2012). His *Safety-I and Safety-II* contrasted two views of safety (Hollnagel 2014). Safety-I treats safety as the absence of failures and focuses on finding and fixing what goes wrong. Safety-II treats safety as the presence of successful adaptation and focuses on understanding why things usually go right. Braithwaite, Wears and Hollnagel carried the argument into health care, proposing that patient safety be turned on its head by moving from Safety-I to Safety-II (Braithwaite et al. 2015). The

founding volume of resilience engineering, edited by Hollnagel, Woods and Leveson, had set out the field's concepts some years before (Hollnagel et al. 2006).

Woods's 2015 paper clarified what resilience engineering means by resilience, and the clarification matters for this book (Woods 2015). He distinguished four concepts that go by the same name: rebound, the capacity to recover from a disruption; robustness, the capacity to absorb disturbances; graceful extensibility, the capacity to stretch when events push a system beyond its normal boundaries; and sustained adaptability, the capacity to keep adapting as conditions change over long periods. He argued that the last two are the field's real subject. This usage differs from the ecological sense of resilience that Chapter 18 introduced through Holling's work, which concerns a system's capacity to persist and absorb change while keeping its relationships intact. The two literatures share a word and some intuitions, and a practitioner reading across them should check which meaning an author intends.

Dekker's *Drift into Failure* brought complexity ideas into the same conversation, arguing that accidents emerge from gradual, normal adaptation, a drift that no one notices because each step seems reasonable, and not from broken parts (Dekker 2010). With Cilliers and Hofmeyr he drew out what complexity theory implies for safety investigations, arguing against root-cause and linear attribution (Dekker et al. 2011). Cook's short document "How Complex Systems Fail", whose site dates it to 1998-2000, distilled eighteen propositions from clinical safety experience, among them that "post-accident attribution to a root cause is fundamentally wrong" and that "safety is a characteristic of systems" (Cook, n.d.). The propositions are persuasive, and it is easy to treat them as findings. They are an expert synthesis, drawn from experience and argument, and they should be read as expert propositions to be tested.

By the end of the 2010s, then, safety science had an unusually rich set of systemic methods, each with its own theory of how accidents happen and its own procedure for analysing them. The natural next question was whether the methods agreed.

One gorge, three methods

Salmon, Cornelissen and Trotter's comparison of AcciMap, HFACS and STAMP is the study that opened this chapter, and its design deserves a closer look because later comparisons followed its pattern (Salmon et al. 2012). The researchers took a single tragedy, the Mangatepopo gorge drownings, and had separate analysts apply each of the three methods to it. AcciMap, derived from Rasmussen's framework, lays out contributing factors across the levels of the sociotechnical system. HFACS, the Human Factors Analysis and Classification System, classifies contributing factors using a fixed taxonomy of failure types. STAMP models the control structure and identifies where control of safety constraints was inadequate. The outputs of the three analyses were then compared and contrasted.

The comparison revealed significant differences across the three methods. The researchers judged AcciMap and STAMP likely to be more comprehensive, capturing a wider range of contributing factors across the system. They judged HFACS, because of its taxonomy, likely to be more reliable, meaning that different analysts using it would be more likely to reach similar conclusions, and more useful for aggregating many cases to find patterns. They recommended a modified AcciMap that would add domain-specific failure taxonomies, combining the comprehensiveness of one approach with some of the reliability of another. The comparison was qualitative; it did not produce overlap statistics, and it should not be summarised with a percentage.

What the study shows is a trade-off. The methods that see more are also the ones that leave more to the analyst's judgment, and the method that constrains judgment with a taxonomy sees less. Better methods are unlikely to eliminate that trade-off, because it reflects a real tension between capturing the particular features of an accident and producing findings that different analysts would agree on and that can be compared across many accidents.

The trade-off has a practical meaning that is worth spelling out. Suppose, hypothetically, that a transport regulator wants to learn from several hundred minor incidents a year. A taxonomy-based method suits that

purpose, because its fixed categories let the regulator count how often each kind of failure recurs and see whether the pattern changes after a new rule comes in. The price is that factors which do not fit the taxonomy tend to disappear from view. Now suppose the same regulator is investigating a single major accident whose causes seem to run through the design of a system, the commercial pressures on its operators and the regulator's own oversight. A comprehensive method suits that purpose better, because it can follow the causes wherever they lead. The price is that a second team using the same method might follow them somewhere else. Neither choice is wrong. What the Mangatepopo comparison makes clear is that the choice of method is a choice about what the investigation will be able to see, and that it should be made deliberately and reported openly.

Grayrigg, Sewol and a collision-avoidance system

Three later studies extended the comparison in different directions, and each adds something the Mangatepopo study could not show.

Underwood and Waterson analysed the Grayrigg train derailment using three methods: the ATSB model, which is based on the Swiss cheese model, together with AcciMap and STAMP (Underwood and Waterson 2014). All three methods applied systems thinking to the derailment. STAMP embodied systems theory most fully. The Swiss cheese model, however, remained viable: it produced a useful systemic analysis, and the comparison did not support discarding it in favour of the newer methods. That finding matters because the popular story of safety science often presents the Swiss cheese model as a linear relic that systemic methods have superseded. The Grayrigg comparison suggests that the older model, applied well, still does much of the work.

Goncalves Filho, Jun and Waterson examined the Sewol ferry accident through four independent analyses using the two methods, AcciMap and STAMP (Goncalves Filho et al. 2019). This design allowed them to measure agreement both within and between methods. Analysts using the same method agreed to a moderate extent: the overlap among analysts using STAMP was 65 per cent, and among analysts using AcciMap 38 per cent.

The overlap between methods was only about 8 per cent. Two teams using different methods on the same accident identified almost entirely different sets of contributing factors.

That 8 per cent figure is the most striking number in this chapter, and it needs careful reading. Low overlap is compatible with both methods being sound. What the figure demonstrates is that the methods direct attention to different things, and that the account of an accident an investigation produces depends heavily on which method it used. If the purpose of an investigation is to find the causes of an accident, low between-method agreement raises a question about reliability that the field has not answered: which of the two accounts should guide prevention, and on what grounds? The within-method figures add a second caution. Even analysts using the same method disagreed substantially, and more so with AcciMap than with STAMP.

The third study moved from accident analysis to hazard analysis in design. Sulaman and colleagues applied FMEA, a long-established method for analysing how components can fail and with what effects, and STPA to the same forward collision-avoidance system (Sulaman et al. 2019). The two methods delivered similar results. That is a sobering finding for claims that STPA is superior to older, component-focused methods. In this case, at least, the systemic method did not find hazards that the traditional method missed.

The Sulaman study differs from the others in an important way. It concerns hazard analysis during design, before anything has gone wrong, and not the explanation of an accident after the fact. That makes it relevant to one of the strongest claims made for STPA, that its control-structure view lets designers anticipate hazards arising from interactions that component-focused methods overlook. One case study cannot settle that claim. It does show that the claim needs comparative evidence, and that in at least one engineered system the traditional method held its own.

Read together, the studies support a clear conclusion. Systemic accident models find different causes from linear and Swiss cheese models, and from each other, and the comparison studies document the differences well. They

do not show that the systemic methods find better causes, in the sense of causes that are more valid or that lead to more effective prevention. I'd grade the claim that systemic accident models find different or additional causes C: the methods clearly produce different outputs, as several independent comparisons show, but the evidence comes from a handful of single-accident case comparisons, the older Swiss cheese model held up in one of them, a traditional method matched the systemic one in another, and low between-method overlap raises reliability questions that the claim of superiority cannot answer.

Where the factors land and who uses the methods

A broader look at the literature confirms and complicates this picture. Hulme and colleagues systematically reviewed 73 published applications of systemic accident analysis methods between 1990 and 2018: 20 using AcciMap, 43 using HFACS, 6 using STAMP or CAST and 4 using FRAM (Hulme et al. 2019). Across these applications, the contributory factors identified still clustered at the lower levels of the system, at the sharp end, close to the operators. Methods designed to draw attention to the blunt end, to management, regulation and government, were in practice producing findings concentrated near the operators. The distribution also shows how uneven the use of the methods is. HFACS, the taxonomy-based method, accounted for more applications than the other three combined.

Why the factors cluster at the sharp end is not settled by the review. It may be that information about the sharp end is easier to obtain after an accident, that investigators are trained to look there, or that the methods, as applied, do not push hard enough upward. Whatever the reason, the finding is a check on the claim that adopting a systemic method automatically shifts an investigation toward the organisational causes that the systems turn emphasised.

The methods are also used less than the research literature might suggest. Underwood and Waterson interviewed 42 safety experts in ten countries and found that systemic accident analysis methods were little used in practice, a gap between research and practice that the authors documented

and examined (Underwood and Waterson 2013). Waterson and colleagues, reviewing sociotechnical systems methods for workplace safety through six case studies, set out the methodological challenges facing an effective science of sociotechnical systems and safety (Waterson et al. 2015).

The newer approaches have similar evidential limits. A systematic review of resilient health care identified 36 studies and found them largely qualitative, with measurement tools derived from predefined concepts (Iflaifel et al. 2020). A literature review of FRAM applications found growing use and variable rigour (Patriarca et al. 2020). Neither review found outcome studies, and neither was designed to.

Credit and limit

The case for crediting safety science is strong, and it should be made plainly. The systems turn changed what the field's own methods direct investigators to look for. The older picture, as the systemic writers described it, pinned the accident on whoever was nearest to it and aimed the fix at that individual. The methods reviewed in this chapter instead ask about design, latent conditions, production pressure, the normalisation of deviance and the control structure above the operator. That change is visible across the research literature, even if practice has followed more slowly, and it is a real achievement. Even the finding that contributory factors still cluster at the sharp end is a finding the field produced about itself.

The second ground for credit is rarer. The field ran head-to-head comparisons of its own tools and put the outcomes in print, including outcomes that did its newest methods no favours. The Mangatepopo, Grayrigg, Sewol and collision-avoidance comparisons are exactly the kind of study this book has looked for, and mostly failed to find, in other parts of the systems field. No comparable set of studies has compared causal loop diagramming with other ways of mapping a problem, or different leverage frameworks on the same case, or different soft systems methods on the same situation. Safety science has done more self-examination than most of the field, and its researchers deserve full credit for that.

The limit is equally plain. No controlled or longitudinal study has shown that adopting STAMP, STPA, FRAM or Safety-II reduces accidents or harm. The claims for these approaches rest on theory, case analyses and comparisons of what different methods find, and a comparison of what methods find is not evidence about what happens to accident rates when organisations adopt them. I'd grade the claim that STAMP, STPA and resilience engineering improve safety outcomes D, because no outcome studies of either kind were found, and the reviews of resilient health care and FRAM describe a literature that is growing in volume without yet testing effects on harm.

The absence of outcome evidence has understandable causes, and they are worth stating because they apply to any attempt to evaluate a systemic method. Serious accidents are rare in most well-run systems, so an organisation that adopts a new method may need many years before a change in its accident rate could be distinguished from chance. Organisations that adopt new methods usually change other things at the same time, which makes it hard to attribute any improvement to the method. And the methods themselves change what gets recorded as an incident, so the measure used to judge them can shift when they are adopted. None of these problems is insurmountable. A health system or transport network could introduce a method in stages across sites, in the manner of the stepped-wedge trial that Chapter 17 describes for public health, and compare trends in incidents and harm between sites that have and have not yet adopted it. Leading indicators of the kind Leveson proposed could serve as intermediate outcomes while the rarer serious events accumulate (Leveson 2015). The field has the concepts it would need to test its claims. What it lacks so far is the studies.

It would be easy to read the combination of these judgments as a verdict against safety science. I think it points the other way. The field did what this book argues systems thinking should do. It used systems ideas as a way of seeing, which changed the questions investigators ask. It treated its methods as claims to be tested, which revealed that they produce different accounts of the same event. And it has been candid about the gap between what its methods change in an investigation and what they have been shown to

change in the world. The leverage illusion appears in safety science only where advocates claim that adopting a systemic method will, by itself, make systems safer. The research literature mostly does not make that claim, and where the popular literature does, the field's own comparisons supply the correction.

A two-method incident review protocol

The comparative studies suggest a practical protocol for any organisation that investigates incidents. Its central idea is simple: analyse important incidents with two different methods, compare the outputs, and report how much they overlap. The protocol draws directly on the design of the Sewol and Mangatepopo comparisons, adapted for routine use. Its evidence base is C, the same as for the claim that different methods find different causes; there is no evidence yet that organisations using it prevent more incidents, and I would present it as a way of making an investigation's dependence on method visible.

1. **Choose two methods that differ in approach.** A good pairing combines one method that is comprehensive and judgment-heavy, such as AcciMap or STAMP, with one that is more structured, such as a taxonomy-based method like HFACS or a Swiss cheese-based model. Pairing two similar methods will understate the dependence on method.
2. **Use separate analysts.** Assign each method to a different analyst or team, working from the same evidence file and not conferring until both analyses are complete. This mirrors the separation in the published comparisons and prevents one account from shaping the other.
3. **List contributing factors in a common format.** Ask each team to produce a list of contributing factors, each stated in a sentence and tagged with the system level at which it operates, from the frontline task up to management, the organisation and external regulation.
4. **Match and count.** Have a third person, or both teams together, match factors that describe the same thing. Count the factors unique to each analysis and those shared. A simple overlap measure is the number of

shared factors divided by the number of distinct factors identified by either team. This is a practical measure I suggest for routine use, and it is not the metric used in the published studies.

5. **Report the overlap and the differences.** Include the overlap figure in the investigation report, together with a short account of what each method found that the other did not. Record where each analysis placed its factors by system level, and note whether either concentrated them at the sharp end.
6. **Decide on actions with the differences in view.** Where the two analyses agree, act with more confidence. Where they differ, treat the factors found by only one method as hypotheses that need further evidence before they justify major changes.

A worked example shows how the numbers might look. Consider a hypothetical hospital investigating a serious medication error. The team using a STAMP-style analysis identifies 18 contributing factors, including gaps in the pharmacy's control over prescribing software updates and a staffing policy that left one ward without a senior nurse at night. The team using a taxonomy-based analysis identifies 12, mostly concerning the nurse's workload, the labelling of the drugs and the supervision on the shift. When the factors are matched, five describe the same thing. The distinct factors number 25, so the overlap is five in 25, or 20 per cent. The report states that figure, lists the factors each method found alone, and notes that the taxonomy-based analysis placed most of its factors at the ward level while the STAMP-style analysis placed several at the level of hospital policy and software governance. The numbers are invented, but the value of the exercise is real in principle: it shows the hospital's board how much of its picture of the incident depends on the method used, and it stops a single account from being mistaken for the whole truth.

Two cautions apply. First, low overlap is not a sign that the investigation failed. The Sewol study found low between-method overlap among careful, independent analysts, and a routine review should expect the same. Second, high overlap does not prove that the shared factors are the true causes. Agreement between methods increases confidence, but the test of any

contributing factor is whether addressing it reduces the recurrence of similar incidents, and that test can only be run by following what happens after the recommendations are implemented. An organisation that records its overlap figures and then tracks the recurrence of incidents over several years would be generating exactly the outcome evidence that the field still lacks.

Chapter 20. Levers Without Illusions: Capability Traps and a Practitioner's Code

Few management stories are as uncomfortable as the one in which everything works. The firm Analog Devices once ran a total quality management programme that by its own measures was a triumph. Manufacturing yield doubled, cycle time was halved and defects fell tenfold. The company's financial performance then got worse. Sterman, Repenning and Kofman, who studied the case and built a model of it, called the result a paradox of organisational improvement: a programme that succeeded on every operational measure was followed by financial decline, and the decline was connected to the success (Sterman et al. 1997).

This chapter uses that case, and a second one that ended better, to pull the book's argument together. It then sets out a short code for practitioners, drawn from the evidence of the previous nineteen chapters, and it closes where the book began, with the bathtub.

A quality programme that succeeded

The Analog Devices case is cautionary for two reasons, and the second is easy to miss. The first is the paradox itself. Sterman and his colleagues explained the decline through unanticipated feedback between improvement, capacity and layoffs (Sterman et al. 1997). In outline, their account runs as follows. Improvement programmes raise productivity, and higher productivity creates capacity. What an organisation does with capacity it does not immediately need feeds back on the programme that created it. If the freed capacity cannot be absorbed by growth, pressure builds to cut costs, and cutting staff in the wake of an improvement drive tells the

workforce something about what improvement is for. The model showed how the same success that the programme's measures recorded could set in motion feedback that its measures did not track.

The second reason concerns how the case was read at the time. Garvin's 1993 article on building a learning organisation named Analog Devices among its examples (Garvin 1993). Garvin's argument was that learning organisations needed clearer meaning, management and measurement, and the company's quality work fitted that picture well. I raise this without any wish to criticise Garvin, whose subject was how learning organisations should be built, because the juxtaposition shows how a case can serve as a positive example of one framework while a feedback analysis of the same case finds a trap. Chapter 13 examined the learning-organisation literature and its evidence; the point here is narrower. An organisation that is learning, by the measures in use, can still be caught by feedback that those measures do not capture.

For this book's argument, Analog Devices is a case of the leverage illusion at the level of the programme. The quality programme was a lever, and it moved exactly what it was designed to move. The illusion lay in the assumption that moving those variables would move the whole system in the same direction. The structure that determined the financial outcome included capacity, demand, cost pressure and the workforce's response to layoffs, and none of that was in view when the lever was pulled.

Working harder, working smarter

Repenning and Sterman generalised the lesson in two papers that developed the idea of capability traps. The first, whose title observes that nobody ever gets credit for fixing problems that never happened, used field studies and simple models to explain why process improvement usually fails (Repenning and Sterman 2001). Its central mechanism is a choice that managers face repeatedly under pressure. When performance falls short, they can ask people to work harder, which raises output quickly, or to work smarter, by investing time in improving the process, which lowers output at first and raises capability only later. Working harder wins in the short run. But time

spent working harder is time not spent improving, so capability erodes, performance falls further short, and the pressure to work harder grows. The organisation becomes trapped in a pattern where the response that looks sensible at each step deepens the problem over time.

The second paper, an inductive study of one firm's improvement efforts, added a psychological element (Repenning and Sterman 2002). Managers who saw performance falling short tended to make dispositional attributions, blaming the workers rather than the situation. The study showed how those attributions interacted with the feedback structure described above to produce capability traps that confirmed themselves: the managers' explanation led to actions whose results appeared to bear the explanation out.

Oliva and Sterman found a related pattern in a bank's lending operation, where they combined field study with a system dynamics model (Oliva and Sterman 2001). Work pressure and management practice pushed service providers toward cutting corners, and quality eroded gradually over time. The pattern resembles the normalisation of deviance that Vaughan found at NASA, described in Chapter 19. The mechanisms differ, but both show how a stock of capability or quality can drain away through flows too small to notice in any single period.

Spear and Bowen's study of Toyota offers a contrast (Spear and Bowen 1999). Over four years, they studied more than 40 plants and proposed four tacit rules behind Toyota's system. The study is a case study, with the limits of one, and it does not test whether the rules cause Toyota's performance. Its relevance here is the contrast between tacit rules that shape how work is done every day and improvement run as a programme, with a launch date and a success metric that can be met and then left behind. I suspect that contrast matters for capability traps, because a programme can be declared finished while a rule of daily work cannot, but no study in this book's sources tests that suspicion.

I'd grade the evidence for capability traps C. The field studies are rich, the models are theoretically coherent and formally specified, and the pattern recurs across the firms studied, but the evidence comes from a small number

of cases, often analysed by the same research group, and nobody has tested how general the trap is in a large sample. What makes the idea valuable despite that grade is that it is specific. It names a feedback mechanism, predicts what an organisation caught in it will look like, and suggests what would have to change, which is more than most organisational frameworks offer.

A review that reached policy

The book's positive closing case comes from child protection in England. In 2011 Munro's review of the child protection system published its final report, subtitled "A Child-Centred System" (Munro 2011). The review was systems-informed, and its central argument was that compliance-driven bureaucracy had displaced professional judgment. In the review's analysis, the requirements accumulated through earlier efforts to improve the system had come to crowd out the professional judgment they were meant to support. The government published a formal response in July 2011.

Lane, Munro and Husemann later gave a behind-the-scenes account of the methods used (Lane et al. 2016). The review blended several systems approaches, including causal analysis in the style of system dynamics alongside other systems ideas, to understand how the parts of the child protection system interacted. Their account is useful for practitioners because it shows systems methods being used in the service of a real policy review, under the constraints of a national review.

The Munro review deserves credit as one of the clearest cases of systems analysis reaching national policy. It is worth being exact about what that credit covers. The review's analysis received a formal government response in the same year as its final report. That is a real achievement, and one that most of the systems applications examined in this book cannot claim. The sources used in this book do not show, and I do not claim, that the reforms that followed improved outcomes for children. Reaching policy and improving outcomes are different achievements, and the review has demonstrated the first. It is also worth noticing the direction of its argument. The review used systems analysis to argue for restoring room for

professional judgment, which is close to the opposite of the leverage illusion. Its analysis treated the accumulated requirements of earlier reforms as part of the problem to be understood, and declined to add one more lever to the pile.

The Munro case also invites comparison with a very different strand of public-sector systems thinking. Seddon's polemic on the failure of the reform regime in UK public services applied ideas from Deming and Toyota, including the concept of failure demand, and argued for a systems approach to service design (Seddon 2008). Its critique of the reform regime overlaps with Munro's critique of compliance-driven bureaucracy, but its evidence consists of case anecdotes rather than systematic analysis. The overlap in conclusion and difference in method is instructive: two systems-informed critiques reached similar views, and one of them was built in a way that could stand up to scrutiny in a national review.

What the evidence delivered

The two cases frame the book's verdict, and it can now be stated directly.

The diagnosis is strong. Chapters 1 to 3 showed that people, including highly educated people and professionals, systematically misjudge accumulation, delay and feedback, and that the failure appears across laboratories, countries and populations. Within that group, Chapter 2 found that presentation changes how large the failure is, so the finding is robust but conditional. Chapter 4 showed a vivid if less rigorously established catalogue of failure in complex simulated worlds, and Chapter 5 showed that unstructured experience teaches less than people expect. Those findings are the field's great contribution, and they belong among the best-established results in behavioural decision research.

The promise is weak. The belief that training people in systems thinking makes their real decisions better, examined in Chapter 14, rests on laboratory proxies and classroom gains; I'd still grade it D, because no field trial or longitudinal study has linked training to the quality of decisions in organisations or policy. The claims that maps change policy (Chapter 17), that leverage points can be ranked (Chapter 12), that archetypes diagnose

real problems (Chapter 13), that Cybersyn demonstrated cybernetic governance (Chapter 16) and that systemic safety methods improve outcomes (Chapter 19) all sit at D. The pattern is consistent. The further a claim moves from the laboratory toward the organisation, the policy or the population, the thinner its evidence becomes.

The lopsidedness has an explanation, and it is uncomfortably close to the capability trap described earlier in this chapter. Laboratory studies of reasoning are relatively cheap and quick, and other teams can replicate them without waiting years for results. Field evaluations of interventions are expensive and slow, and when their results arrive they are often mixed, as the STOPS trial of Chapter 17 was. A field under pressure to show relevance will tend to produce more applications, more maps and more case reports, which is the equivalent of working harder, and fewer rigorous evaluations, which is the equivalent of working smarter. Each application looks like progress, but the stock of tested knowledge about what systems interventions achieve grows slowly, and the gap between promise and evidence widens. I do not want to push the analogy too far, because I know of no study of the systems field's research choices in these terms. It does suggest why the gap has persisted for so long among researchers who understand feedback better than almost anyone.

Between the diagnosis and the promise lie the claims that earn middle grades, and several of them point toward how the field could improve. Group model building reliably improves shared understanding and consensus in the room, as Chapter 11 showed, even though its effects months later are unknown. Ostrom's design principles, in Chapter 18, describe robust common-pool resource institutions across many cases. Psychological safety, in Chapter 13, has a large meta-analytic base. The one rigorous trial of a whole-systems intervention, STOPS in Chapter 17, produced mixed weight outcomes but showed that such trials can be run. And the behaviour-rich COVID-19 models discussed in Chapter 17 gave the field a rare quantitative test of one of its claims against real forecasts.

That last result bears on a debate about model size that practitioners meet constantly. Ghaffarzadegan, Lyneis and Richardson argued that small, transparent system dynamics models can reveal policy resistance and delays to policymakers more cheaply and credibly than large models (Ghaffarzadegan et al. 2011). C-ROADS, a small climate policy simulator used with negotiators under the UN climate convention to link emissions pledges to carbon dioxide concentrations and warming, is a well-documented example (Sterman et al. 2012). The argument is plausible, and practitioner experience supports it. Yet the COVID-19 forecasting study complicates it: the models that forecast best over longer horizons were those that included the right mechanisms, including endogenous behavioural responses, and a simple model with those mechanisms matched the best models up to 20 weeks ahead, which suggests that mechanisms matter more than size (Rahmandad et al. 2022). Rahmandad and Sterman's comparison of agent-based and differential-equation models reaches a compatible conclusion from a different direction: the two types differ on some metrics, sometimes only slightly, and the choice between them should weigh costs against benefits (Rahmandad and Sterman 2008). I'd grade the claim that small, transparent models are more useful to policymakers C: it is plausible and supported by practitioner accounts and examples, but no comparative study has tested it, and the best quantitative evidence available points to mechanism rather than size as the thing that matters.

The field has begun to say much of this about itself. Sterman's agenda for system dynamics at sixty called for more attention to data, more rigour and more integration with other disciplines (Sterman 2018). Saltelli and colleagues' manifesto for responsible modelling asked modellers to mind their assumptions, their hubris, their framing, the consequences of their models and the unknowns they cannot model (Saltelli et al. 2020). Chapter 15 set out the argument, from Oreskes and colleagues, that models of open systems can be evaluated and never fully validated. Those three sources, read together, amount to a programme for a more honest discipline, and the safety science of Chapter 19 shows that parts of the systems world have already begun to carry it out.

A practitioner's code

The evidence supports a short code for anyone who uses, commissions or teaches systems thinking. It has seven points, and each follows from something the book has shown. The code itself is a set of judgments drawn from graded evidence, and I would not claim that following it has been shown to improve outcomes. It is offered as the practice most consistent with what the evidence supports.

Start from the diagnosis. The field's grade-A results all concern one thing: people misjudge accumulation, delay and feedback. Every systems project should begin by asking where those misjudgements are likely to arise in the problem at hand: which stocks are being read off their flows, which delays are being ignored, which supply lines are being forgotten. This is the part of systems thinking that rests on the firmest evidence, and it is often skipped in favour of the more exciting search for levers.

Present dynamics in formats people can read. Chapter 2 showed that the size of the stock-flow failure depends on presentation, and that verbal framing and attention to the global pattern can reduce it. Anyone presenting dynamic information to decision makers should choose formats that make accumulation visible, test whether the audience reads them correctly, and assume that a graph which seems obvious to the analyst will be misread by many of the people it is meant to inform.

Treat maps as hypotheses. A causal loop diagram, a system map or a group model is a hypothesis about structure. Chapter 10 showed what diagrams hide, and Chapter 12 showed that reading a lever off a map is the step with the least evidence behind it. Every map should come with a statement of what evidence would show it to be wrong, and with a plan to gather some of that evidence.

Measure proximal and distal outcomes. Chapter 11 showed that group model building earns a B for its effects on participants' shared understanding and a D for its effects months or years afterwards. The split applies far more widely. A project that measures only whether participants found it useful

has measured the proximal outcome. The distal outcome, whether the decisions, the organisation or the population changed, is what the project was for, and it should be measured too, even when the measurement is hard.

Grade your own claims. The rubric used in this book, explained in the front matter and set out in full in Appendix A, can be applied by anyone to any claim. Before presenting a systems finding, grade it: what is the design, what is the sample, has it been replicated, and is the outcome proximal or distal? A practitioner who grades their own claims will make fewer of them, and the ones they make will be more credible.

Use leverage points and archetypes as prompts for questions. Chapter 12 showed that Meadows's list is an explicitly tentative heuristic, and Chapter 13 showed that archetypes are widely taught and rarely tested. Both are useful for generating questions about where change might be possible. Neither is evidence that change at a given point will work, and a proposal that rests on either should say what separate evidence supports it.

Build the evaluation in from day one. Chapter 17 showed that serious evaluation of a systems intervention is possible, and that it has to be designed before the intervention begins. The same holds for a modelling project, a group model building exercise or a new incident review method. Decide at the start what outcome would count as success, how it will be measured, what the comparison will be, and when the result will be known.

For use in practice, the code can be reduced to a single page, with each point linked to the part of the book and the appendix that supports it:

1. Start from the diagnosis: find the likely stock, delay and feedback errors first (Part I; Appendix A, grade A claims; Appendix D, cards 1 and 3).
2. Present dynamics in readable formats, and test that they are read correctly (Chapter 2; Appendix D, card 2).
3. Treat every map and model as a hypothesis, with stated tests (Chapters 10, 12 and 15; Appendix D, cards 10 and 15).
4. Measure proximal and distal outcomes separately (Chapters 11 and 17; Appendix A).

5. Grade your own claims using the design, sample, replication and outcome questions (Appendix A).
6. Use leverage points and archetypes to generate questions, and seek separate evidence for any lever (Chapters 12 and 13; Appendix D, cards 12 and 13).
7. Design the evaluation before the intervention (Chapter 17; Appendix D, card 17).

A hypothetical example shows how the code works on a real-looking proposal. Suppose a regional government is asked to fund a systems programme to reduce hospital readmissions. The proposal includes a causal map built in workshops with clinicians and managers, identifies discharge planning as the leverage point, and promises a large fall in readmissions within two years. Running the code takes an afternoon. The first point asks where the stock-flow errors lie, and it turns out that the proposal treats the monthly readmission rate as if it would respond immediately to a change in discharge planning, ignoring the delay while a changed practice spreads across wards. The second asks how the programme will report progress to the board, and suggests showing the stock of patients at risk alongside the monthly flow. The third asks what evidence would show the map is wrong, which leads to a small audit of recent readmissions to see whether discharge failures are actually common among them. The fourth separates the workshop's success, which the proposal documents well, from the readmission outcome, which it does not yet measure. The fifth grades the central claim, that improving discharge planning will cut readmissions sharply, at D on present evidence, because it rests on the map and on anecdote. The sixth notes that calling discharge planning the leverage point is a prompt for inquiry and asks for separate evidence. The seventh requires a comparison, perhaps by introducing the change ward by ward, before the programme starts. None of this kills the proposal. It turns a promise into a test, which is what the evidence in this book suggests a systems proposal should be.

Back to the bathtub

Chapter 1 asked you to write down your answers to two short problems before reading any further. One of them asked what happens to the carbon dioxide in the atmosphere if emissions are held steady at a rate above the rate at which carbon dioxide is removed. If you still have your answers, look at them now. The stock keeps rising for as long as the inflow exceeds the outflow, and it stops rising only when emissions fall to the rate of removal. Many careful, educated people answer otherwise, and Chapter 1 explained why that error is so common and so persistent.

Your answers are a small piece of evidence about the first half of the leverage illusion. If you answered correctly, you may have done so because the problem was framed in a way that made the stock visible, or because you have practised this kind of reasoning. If you answered incorrectly, you have joined a large and distinguished company. Either way, the exercise shows something that no diagram could: reading even a one-stock system is harder than it feels, and the feeling of having read it correctly is no guarantee that you did.

The second half of the illusion, the belief that seeing the structure hands you a reliable lever, is where the book has found the evidence thinnest. It is also where the field's most famous essay was most cautious. Chapter 12 showed that Meadows herself called her list tentative and doubted that anyone had found an easy formula for locating leverage points, and it quoted the closing line of her essay, in which mastery of systems turns out to lie more in letting go than in pushing (Meadows 1999). I read that ending as an invitation to hold our levers more lightly: to treat every claim about where to push as a hypothesis, to watch what happens when we push, and to be willing to learn that we were wrong.

That is the standard systems thinking sets for itself, and it is the standard this book has applied. A systems thinker is supposed to trace a loop all the way back to where it started, and for a method that means tracing what it claimed through to what was later measured. Where the loop closes, as it has in the laboratory work on accumulation, in Ostrom's comparative studies and in safety science's comparisons of its own methods, systems

thinking has earned its authority. Where it remains open, the practitioner's job is to close it, by grading the claims that are made and measuring the outcomes that follow. When you next meet a claim that a system has a lever, take out the pencil and the scrap of paper that Chapter 1 asked for, sketch the stocks and the delays, and then ask the question this book has asked of every lever it met: what evidence would show that pushing here does what we hope?

Appendices

Appendix A. Evidence Grades Table

This appendix gathers in one place every graded claim in the book. There are 38 of them, grouped in seven clusters by subject. For each claim the table gives the grade, the sources that most strongly support it, the sources that most strongly contradict or qualify it, and the chapter or chapters where it is discussed. The chapters give the full reasoning; the table is a map back to them.

The grading rubric

Each claim is graded on the four-point scale introduced in the front matter.

- **A (strong).** Replicated experimental or quantitative findings from several independent teams, with known boundary conditions. The claim can be stated without heavy hedging.
- **B (moderate).** Consistent evidence from several studies or a good meta-analysis, but with limits of design (correlational, laboratory-only or self-reported) or rival explanations that have not been ruled out.
- **C (limited or mixed).** Some supportive studies, but with small samples, single cases, before-and-after designs without controls, or conflicting results. A C claim is best treated as a plausible hypothesis.
- **D (weak, untested or contradicted).** Evidence consisting mainly of theory, practitioner testimony or anecdote; no outcome studies; or the best evidence points the other way. Where a D means that the evidence contradicts the claim, the table says so.

Two conventions matter for reading the table. First, each grade judges the claim as it is popularly stated. A narrower version of the same claim may deserve a higher grade, and where it does, both grades are given: group

model building, for example, earns a B for its effects on participants' shared understanding and a D for its effects on later organisational or policy outcomes. Second, a grade describes the state of the evidence, not the importance or the plausibility of the idea. Several D-graded claims are reasonable ideas that nobody has yet tested.

In the table, "Ch." gives the chapter that owns the discussion first, followed by any chapter that returns to it. Citations are shortened; the full references are in the References section at the end of the book.

The 38 claims

#	Claim	Grade	Main support	Main contradiction or qualification	Ch.
	Cluster I: How people reason about dynamic systems				
1	People systematically misjudge stock-flow (accumulation) relationships in graph and table formats	A	Sweeney and Sterman 2000; Cronin, Gonzalez and Sterman 2009; Abdel-Hamid et al. 2014	Fischer, Degen and Funke 2015 (verbal formats); Fischer and Gonzalez 2016	1, 2
2	The correlation heuristic is the main mechanism	B	Cronin et al. 2009; Korzilius et al. 2014	Cronin and Gonzalez 2007; Fischer and Gonzalez 2016 (rival mechanisms)	2
3	People misperceive feedback and delays, producing oscillation	A	Sterman 1989a; Diehl and Sterman 1995; Croson and Donohue 2006	Wu and Katok 2006; Steckel, Gupta and Banerji 2004 (remedies help partly)	3
4	The bullwhip effect is partly behavioural	A	Croson and Donohue 2006; Sterman and Dogan 2015	Lee, Padmanabhan and Whang 1997 (operational causes also produce it)	3
5	Experience and outcome feedback alone do not fix misperceptions	B	Paich and Sterman 1993; Moxnes 2004	Kopainsky and Sawicka 2011; Gonzalez and Wong 2011 (structured experience helps)	5
6	Sharing information across a supply chain cures oscillation	C	Croson and Donohue 2006 (helps somewhat)	Steckel et al. 2004; Wu and Katok 2006	3
7	Resource overuse arises from misperception even without a commons dilemma	B	Moxnes 1998; Moxnes 2000	None found; replication outside one programme limited	5, 18
8	Expertise and higher education do not protect against stock-flow errors	B	Abdel-Hamid et al. 2014; Brunstein, Gonzalez and Kanter 2010	Sterman 2010 (targeted training helps)	1, 2
9	Short system dynamics training improves stock-flow reasoning	C	Sterman 2010; Kainz and Ossimitz 2002	No randomised controls; a minority still reason correlationally	5, 14

#	Claim	Grade	Main support	Main contradiction or qualification	Ch.
10	Systems-thinking training improves real-world decisions	D	Capelo and Dias 2009 (laboratory proxy)	Maani and Maharaj 2004; no field trial exists	14, 20
11	Dörner's microworlds show a stable taxonomy of failure	C	Dörner et al. 1983; Dörner 1996	Kluge 2008; Greiff, Wüstenberg and Funke 2012 (reliability)	4
12	Complex problem solving is distinct from intelligence	B overlapping; D fully distinct	Wüstenberg, Greiff and Funke 2012	Stadler et al. 2015 (meta-analysis); Kretzschmar et al. 2016	4, 14
13	Systems thinking can be measured as a coherent skill	C	Plate 2010; Mehren et al. 2018	Dugan et al. 2022 (fragmented measures)	14
	Cluster II: Modelling and participatory methods				
14	Group model building improves shared understanding and consensus	B proximal; D downstream	Rouwette, Vennix and Mullekom 2002; Scott, Cavana and Cameron 2016	Mostly self-report and single cases	11
15	System dynamics modelling improves health policy outcomes	D	Homer and Hirsch 2006; Milstein et al. 2011 (projections)	Atkinson et al. 2015 (no analytic studies); Darabi and Hosseinichimeh 2020	17
16	Small, transparent models serve policymakers better than large ones	C	Ghaffarzadegan, Lyneis and Richardson 2011; Sterman et al. 2012	Rahmandad, Xu and Ghaffarzadegan 2022 (mechanisms matter more than size)	20
17	Models with endogenous behavioural feedback forecast better	C, moving toward B	Rahmandad et al. 2022; Rahmandad, Lim and Sterman 2021	One epidemic, one team; Saltelli et al. 2020	17, 20
18	System dynamics models can be validated in a meaningful sense	C	Barlas 1996; Sterman 2000	Oreskes, Shrader-Frechette and Belitz 1994; Castro 2012	15, 20
	Cluster III: Grand models and their critics				
19	Social systems are counter-intuitive; intuitive policies often backfire	C	Forrester 1971a; Sterman 2006	Kadanoff 1971; Lee 1973; no base rate exists	8, 12
20	World3's standard run has tracked observed data since 1972	C	Turner 2008 (a limited "degree of validation"); Herrington 2021; Nebel et al. 2024	Castro 2012; Oreskes et al. 1994 (aggregate tracking is not a structural test)	15
21	Nordhaus was right that *World Dynamics* lacked empirical grounding	B	Nordhaus 1973; Solow 1973	Forrester, Low and Mass 1974 (reply)	15
22	Cybersyn showed that cybernetic real-time governance works	D	Medina 2006; Medina 2011 (historical case)	Scott 1998; Jackson 1988; no outcome evidence either way	16
	Cluster IV: Organisations				
23	Learning organisations outperform	C	Goh, Elliott and Quon 2012 (meta-analysis)	Grieves 2008; causality unresolved	13

#	Claim	Grade	Main support	Main contradiction or qualification	Ch.
24	Team psychological safety predicts learning and performance	B	Edmondson 1999; Frazier et al. 2017	Mostly correlational	13
25	Capability traps undermine improvement efforts	C	Repenning and Sterman 2001; Sterman, Repenning and Kofman 1997	No large-sample tests	20
26	System archetypes diagnose real organisational problems	D	Senge 1990; Wolstenholme 2003	Lane 2008; Richardson 1986; no comparative test	13
27	Soft systems methodology and problem structuring methods are effective	C use; D comparative	Mingers and Taylor 1992; Mingers and Rosenhead 2004	Self-selected users; no comparative designs	9
28	Cynefin improves decisions	D	Kurtz and Snowden 2003	No empirical tests found	13
	Cluster V: Policy and public health				
29	Systems thinking in public health is expanding but mostly conceptual	B	Carey et al. 2015; Chughtai and Blanchet 2017; Rusoja et al. 2018	None found	17
30	Whole-systems community interventions reduce childhood obesity	C	Allender et al. 2021; Bagnall et al. 2019	Allender et al. 2021 (BMI fell, then rose); weak designs elsewhere	17
31	System maps such as the Foresight map change policy	D	Butland et al. 2007; Finegood, Merth and Rutter 2010	Johnston, Matteson and Finegood 2014; Public Health England 2019	17
32	Complex interventions require new evaluation methods	C	Rutter et al. 2017; Skivington et al. 2021	Allender et al. 2021 shows trials are feasible	17
	Cluster VI: Commons, ecology and resilience				
33	Ostrom's design principles characterise robust common-pool resource institutions	B for common-pool resources; D general	Ostrom 1990; Cox, Arnold and Villamayor Tomás 2010; Baggio et al. 2016	Baggio et al. 2016 (context matters); Wilson, Ostrom and Cox 2013 (extension theoretical)	18
34	The tragedy of the commons is inevitable without privatisation or state control	D (contradicted)	Hardin 1968	Ostrom 1990; Ostrom, Walker and Gardner 1992; Cox et al. 2010	18
35	Early-warning signals can anticipate critical transitions	C	Scheffer et al. 2009; Scheffer et al. 2001	Boettiger and Hastings 2012 (false alarms and misses)	18
36	Leverage points can be ranked	D	Meadows 1999 (explicitly tentative)	Dorninger et al. 2020; Johnston et al. 2014 (a coding frame, not a test)	12
	Cluster VII: Safety science				
37	Systemic accident models find more or different causes	C	Salmon, Cornelissen and Trotter 2012; Underwood and Waterson 2014	Goncalves Filho, Jun and Waterson 2019 (about 8% overlap between methods); Sulaman et al. 2019	19

#	Claim	Grade	Main support	Main contradiction or qualification	Ch.
38	STAMP/STPA and Safety-II improve safety outcomes	D	Leveson 2011; Hollnagel 2014	Iflaifel et al. 2020; no outcome studies	19

The pattern in the table

Read down the grade column and a pattern appears that the book's argument rests on. The A grades cluster in the first cluster, in the cognitive laboratory, where people repeatedly fail to reason correctly about accumulation, delay and feedback. The D grades cluster where the field makes its practical promises: that training transfers to real decisions, that models improve policy outcomes, that leverage points can be ranked, that maps change policy, that archetypes diagnose problems and that new safety methods reduce harm. The B grades in between mark the places where careful work has produced solid, if bounded, knowledge: Ostrom's principles for common-pool resources, psychological safety, the proximal benefits of group model building, the growth of systems work in public health and Moxnes's misperception experiments. The field's diagnosis of human error in dynamic systems is far better evidenced than its cure.

Grading a new claim yourself

The grades in this book will date as new studies appear, and you will meet claims it does not cover. The four questions below are enough to place most claims on the rubric with reasonable confidence.

1. **What is the design?** Is the evidence an experiment with a comparison group, a field study observing what happened, a review that pools many studies, a single case, or an argument? Experiments and good reviews can support an A or B. Single cases and arguments rarely support more than a C, and theory alone is a D however persuasive it is.
2. **What is the sample?** How many people, organisations or cases were studied, and how were they chosen? A striking result from a small or self-selected sample, such as practitioners reporting on their own projects, deserves less weight than a modest result from a large or random one.

3. **Has the finding been replicated independently?** A result reproduced by a different team, in a different place, with a different population, is worth far more than the same result reported several times by one group. Much of the gap between the A and B grades in this table comes down to this question.
4. **Is the outcome proximal or distal?** Did the study measure understanding, attitudes or agreement among participants, or did it measure decisions, health, safety or performance in the world? Proximal outcomes are easier to change and easier to measure, and they often improve when distal ones do not. Where a claim is supported only on proximal outcomes, grade it on those and give the distal version its own, usually lower, grade.

A claim that passes all four questions well is probably an A or B. A claim that fails the first or third is probably a C or D whatever its reputation. Write down the grade with its reason, as the chapters do, so that anyone who disagrees can see exactly where the disagreement lies.

Appendix B. Glossary

The definitions below are short working definitions, written for a reader who wants to check a term quickly. Each is paraphrased from the chapter that introduces the term, and the chapter number follows the definition. Where a term has several meanings, the entry says so.

A to C

Accumulation. The process by which a quantity builds up or drains down over time as the running total of what has flowed in minus what has flowed out. (Chapter 1)

AcciMap. A systemic accident analysis method, based on Rasmussen's account of risk management as a control problem spanning every level of society, that maps the contributing factors of an accident across levels from front-line work up to government and regulation. (Chapter 19)

Archetype (system archetype). One of a small library of recurring feedback structures, such as "fixes that fail" or "shifting the burden", popularised by Senge as patterns that managers can learn to recognise. The book treats them as hypotheses, since their diagnostic value is untested. (Chapter 13)

Autopoiesis. Maturana and Varela's term for self-producing organisation: a system, such as a living cell, that continuously produces the components that constitute it. Its later extension into social theory is disputed. (Chapter 6)

Balancing loop. A feedback loop that counteracts change, pushing a system back toward a goal or limit; also called negative feedback. (Chapter 10)

Boundary critique. The practice, developed by Ulrich in critical systems heuristics, of questioning where the boundary of a system has been drawn and whose interests the choice serves. Ulrich's twelve boundary questions are asked in two modes, what "is" the case and what "ought" to be. (Chapter 9)

Bullwhip effect. The amplification of order fluctuations as they pass up a supply chain, so that small changes in customer demand produce large swings in orders at the factory. It has both operational and behavioural causes. (Chapter 3)

Capability trap. A self-reinforcing pattern in which pressure to meet immediate targets leads people to work harder rather than invest in improving their processes, so that capability erodes and the pressure grows. Repenning and Sterman described it from field studies and models. (Chapter 20)

CAST (Causal Analysis based on Systems Theory). The accident analysis method built on STAMP, used to examine why the control structure around a hazard failed to enforce safety constraints. (Chapter 19)

CATWOE. The mnemonic soft systems methodology uses to test a root definition. Its six entries are customers (those who benefit from or suffer from the system), actors (those who carry out its activities), the transformation (the input it changes into an output), the worldview or *Weltanschauung* that makes the transformation meaningful, owners (those who could stop the system) and environmental constraints (what the system takes as given). (Chapter 9)

Causal loop diagram. A diagram of variables joined by arrows marked with polarities, used to show feedback loops. It hides the difference between stocks and flows. (Chapter 10)

Common-pool resource. A resource, such as a fishery, forest, pasture or irrigation system, from which it is costly to exclude users and in which one user's take reduces what is left for others. (Chapter 18)

Community-based system dynamics. A form of group model building, developed by Hovmand and colleagues, in which community members build system dynamics models of problems affecting them, with the aim of building local capacity as well as insight. (Chapter 11)

Complex intervention. An intervention with several interacting components, delivered in a context that shapes how it works; community obesity programmes are an example. The updated Medical Research Council framework guides their development and evaluation. (Chapter 17)

Complex problem solving (CPS). A research tradition, mainly European, that studies how people explore and control dynamic, opaque tasks, usually through computer simulations. It overlaps substantially with general intelligence without being identical to it. (Chapter 4)

Confidence-building tests. The tests a modeller runs to earn trust in a model, since full validation is out of reach. They include tests of structure (does each relationship match what is known?), of behaviour (does the model reproduce history?) and of robustness (do the conclusions survive changes in uncertain assumptions?). (Chapter 15)

Correlation heuristic. The assumption that a stock follows the same pattern as its inflow, so that when the inflow falls the stock must be falling too. (Chapter 2)

Counterintuitive behaviour. Forrester's claim that complex social systems often respond to policies in ways opposite to what intuition predicts, so that well-meant interventions can worsen the problems they target. (Chapter 8)

Critical slowing down. The tendency of a system approaching a threshold to recover more and more slowly from small disturbances. It shows up in data as rising variance and autocorrelation and is the basis of most proposed early-warning signals. (Chapter 18)

Cynefin. A sense-making framework introduced by Kurtz and Snowden that sorts situations into different kinds of context; a popular version for managers by Snowden and Boone distinguished simple, complicated,

complex and chaotic contexts, plus disorder, each calling for a different response. It has not been tested for whether it improves decisions. (Chapter 13)

D to M

Delay. A lag between a cause and its effect. Delays arise wherever a stock stands between an inflow and an outflow, and their length depends on how large the stock is relative to its flows (Chapter 1). The supply line of goods ordered but not yet received is the delay that drives the oscillations of the beer game (Chapter 3), and causal loop diagrams mark delays on the links where they occur (Chapter 10).

Design principles (Ostrom). Eight conditions that characterise long-lived, self-governed common-pool resource institutions: clear boundaries around the users and the resource; rules that fit local conditions, with users' costs in proportion to their benefits; collective-choice arrangements; accountable monitoring; graduated sanctions; low-cost conflict resolution; recognition of the users' right to organise; and nested layers of governance. Reviews support them well for common-pool resources; their extension to all groups is untested. (Chapter 18)

Deviation-amplifying feedback. Maruyama's term for feedback that magnifies an initial change instead of correcting it, which he argued early cybernetics had neglected; the same thing as a reinforcing loop. (Chapter 6)

Dynamic complexity. The difficulty that arises from feedback, delays and accumulation in a system, as distinct from the number of its parts. (Chapter 3)

Early-warning signal. A statistical pattern, such as rising variance or autocorrelation, proposed as a sign that a system is approaching a critical transition. Such signals are supported in some systems but have shown high rates of false alarms and misses. (Chapter 18)

Emergence. The appearance at one level of organisation of properties that are not found in the parts at the level below, as in Anderson's argument that "more is different". (Chapter 7)

Feedback loop. A closed chain of cause and effect in which a change in one element eventually returns to influence that element. Loops either reinforce change or balance it. (Chapter 6; formal notation in Chapter 10)

Flow. A rate of change that fills or drains a stock, measured per unit of time: litres per minute into a bath, cases per week into a warehouse. (Chapter 1)

FRAM (Functional Resonance Analysis Method). Hollnagel's method for describing a sociotechnical system as a set of linked functions. It asks how the ordinary, day-to-day variation in each function could interact with variation in the others, since those interactions, which Hollnagel calls resonance, can produce outcomes nobody intended. (Chapter 19)

Good regulator theorem. Conant and Ashby's formal result that, under broad conditions, a regulator that is both maximally successful and as simple as possible must be a model of the system it regulates. (Chapter 6)

Group model building. A family of facilitated methods in which a client group builds a system dynamics model or diagram together, using scripted exercises. It reliably improves shared understanding and consensus among participants; its downstream effects are unmeasured. (Chapter 11)

HFACS (Human Factors Analysis and Classification System). A taxonomy-based accident analysis method that sorts contributing factors into predefined categories at several levels. It is judged more reliable, but less comprehensive, than AcciMap or STAMP. (Chapter 19)

Homeostasis. The maintenance of a stable internal condition through balancing feedback, as in body temperature. Ashby's homeostat was a machine built to demonstrate adaptive stability of this kind. (Chapter 6)

Increasing returns. A situation in which the more something is adopted, the more attractive it becomes, so that small early events can lock in one outcome, even an inferior one. Arthur's model of competing technologies is the standard formal treatment. (Chapter 7)

Learning organisation. An organisation that deliberately builds its capacity to learn and adapt, in the sense popularised by Senge. Its link to performance is correlational. (Chapter 13)

Leverage illusion. The belief that once you can see a system's structure, you have a reliable lever to move it. This book's central subject. (Front matter; examined in Chapter 12)

Leverage point. In Meadows's sense, a place in a system where a small shift can produce large changes. Meadows proposed twelve such places in a ranked list that she herself called tentative and whose order she called slithery. (Chapter 12)

MicroDYN. An approach to measuring complex problem solving that assembles many short, formally specified dynamic tasks, designed to overcome the low reliability of single, long simulations. (Chapter 4)

Microworld. A computer simulation of a simplified world, such as a town, a region or a firm, used to study or train decision-making in dynamic situations. (Chapter 4)

N to R

Near-decomposability. Simon's term for the property of systems whose parts interact strongly within themselves and only weakly with one another, so that each part can be understood and managed nearly independently. (Chapter 7)

Net flow. Inflow minus outflow at a given moment. The net flow determines whether a stock is rising or falling, whatever the individual flows are doing. (Chapter 1)

Normalisation of deviance. Vaughan's term, from her study of the Challenger launch decision, for the gradual process by which departures from safe practice come to be accepted as normal within an organisation. (Chapter 19)

Organised complexity. Weaver's term for problems in which many factors are interrelated, which he set apart from problems of simplicity and of disorganised complexity. (Chapter 7)

Planetary boundaries. A framework proposed by Rockström and colleagues that defines nine Earth-system boundaries marking a safe operating space for humanity. (Chapter 15)

Polarity. The sign attached to a causal link in a loop diagram. A positive link means that if the cause increases, the effect is higher than it would otherwise have been, and if the cause decreases, the effect is lower than it would otherwise have been; a negative link means the reverse. The clause "than it would otherwise have been" matters, because a stock can go on rising after its inflow falls, so a positive link does not mean that the two variables always move together. (Chapter 10)

Policy resistance. What happens when an intervention sets off responses the intervener did not foresee: other actors adjust their behaviour and feedback loops push back, so that over time the intended effect is diluted, cancelled or reversed. (Chapter 8)

Problem structuring method. One of a family of methods, including soft systems methodology and cognitive mapping, designed to help groups agree what a problem is before trying to solve it. (Chapter 9)

Reinforcing loop. A feedback loop that amplifies change in whichever direction it starts, producing growth or collapse; also called positive feedback. (Chapter 10)

Requisite variety. Ashby's law that a regulator can control a system only if it commands at least as much variety, as many distinct responses, as the disturbances it must counter. (Chapter 6; applied to organisations in Chapter 16)

Resilience (ecological). Holling's term for the capacity of a system to persist and absorb change while keeping the relationships among its parts intact, as distinct from stability, which is the tendency to return to an equilibrium. (Chapter 18)

Resilience (safety science). In Woods's analysis, a word with four meanings: rebound from disruption, robustness, graceful extensibility when events exceed a system's design, and sustained adaptability over time. (Chapter 19)

Rich picture. In soft systems methodology, a free-form drawing of a problem situation that captures its actors, relationships, conflicts and concerns before any system is defined. (Chapter 9)

Root definition. In soft systems methodology, a concise statement of what a relevant system does, how and why, built with the help of CATWOE. (Chapter 9)

S to Z

Safety-I and Safety-II. Hollnagel's contrast between two views of safety: Safety-I defines it as the absence of failures; Safety-II defines it as the presence of successful adaptation. (Chapter 19)

Second-order cybernetics. The cybernetics of observing systems, associated with von Foerster, which treats the observer as part of the system observed and asks how knowledge is constructed. (Chapter 6)

Soft systems methodology (SSM). Checkland's approach to messy problem situations, in which systems models are treated as devices for structuring debate about purposeful action rather than as descriptions of the world. Its learning cycle runs from rich pictures through root definitions to models of purposeful activity. (Chapter 9)

STAMP (System-Theoretic Accident Model and Processes). Leveson's accident model, which treats accidents as the result of inadequate control over safety constraints within a sociotechnical control structure, rather than as chains of component failures. (Chapter 19)

Stepped-wedge trial. A cluster randomised design in which every cluster eventually receives an intervention but the order of starting is randomised, so that clusters still waiting serve as comparisons for those already under way. (Chapter 17)

Stock. Anything that accumulates and can be measured at a moment in time: water in a bath, inventory in a warehouse, carbon dioxide in the atmosphere, people living with a disease. A stock changes only through its flows. (Chapter 1)

STPA (System-Theoretic Process Analysis). The hazard analysis method built on STAMP, used before an accident to identify how unsafe control actions could arise. (Chapter 19)

Supply-line underweighting. The tendency, documented in the beer game and related experiments, to give too little weight to orders already placed but not yet received, leading to over-ordering and oscillation. (Chapter 3)

Swiss cheese model. Reason's image of accidents as the result of holes in several layers of defence lining up, with latent conditions in the organisation creating holes that front-line errors then pass through. (Chapter 19)

System dynamics. Forrester's method of modelling systems as stocks, flows and feedback loops, simulated over time, on the premise that a system's behaviour arises mainly from its internal feedback structure. (Chapter 8)

Tipping point. A threshold beyond which a small change sets off a large, often self-sustaining shift in a system's state. In climate science, the elements that may behave this way are called tipping elements. (Chapter 15)

Validation versus evaluation of models. The distinction, pressed by Oreskes and colleagues, between proving a model true, which they argued is impossible for models of open natural systems, and evaluating or partially confirming it against evidence, which is possible. The book follows it in preferring the language of confidence-building to that of validation. (Chapter 15)

Viable System Model (VSM). Beer's model of the functions any organisation needs to survive in a changing environment: operations (System 1), coordination among them (System 2), management of the operational whole in the present (System 3), attention to the environment and the future (System 4), and policy and identity (System 5). The model is recursive: each operational unit in System 1 is itself a viable system with its own five functions. (Chapter 16)

Whole-systems approach. In public health, an approach that treats a problem such as obesity as the product of many interacting causes and seeks coordinated action across sectors and levels, usually starting from a shared map of the local system. (Chapter 17)

Wicked problem. Rittel and Webber's term for planning problems that have no definitive formulation and no stopping rule, whose solutions are better or worse rather than true or false, and where every attempted solution has consequences. (Chapter 9)

Appendix C. A Reading Path

The bibliography behind this book runs to several hundred items, and nobody needs to read most of them. The four tracks below are short routes through it. The first is where to start if you use systems thinking in your work and want a sound foundation. The other three follow the book's own lines of argument: the evidence, the lineage and the critics. Each entry says in a sentence why it is on the list, and full references are in the References section.

The practitioner's core

Start here. These eight works together give a working grasp of the main methods and of what stands behind them.

- **Meadows (2008), *Thinking in Systems*.** The most widely read introduction to stocks, flows, feedback and leverage points, written as teaching rather than as evidence. Read it for the concepts, and read Chapter 12 of this book alongside it.
- **Sterman (2000), *Business Dynamics*.** The standard graduate text of system dynamics. Selected chapters on stocks and flows, delays, the beer game and model testing repay the effort even for readers who will never build a model.
- **Sterman (2001), "System Dynamics Modeling: Tools for Learning in a Complex World".** A short, accessible statement of policy resistance and of simulators as tools for learning.
- **Meadows (1999), "Leverage Points: Places to Intervene in a System".** The essay at the centre of this book. Read it for its caveats as much as for its list.

- **Checkland (2000), "Soft systems methodology: a thirty year retrospective".** The originator's own account of the soft tradition, and a necessary counterweight to the idea that there is one system to be modelled.
- **Ostrom (1990), *Governing the Commons*.** The best example in the systems literature of claims about governance built from comparative evidence.
- **Leveson (2011), *Engineering a Safer World*.** The book-length statement of STAMP and its methods, and the clearest account of the systems turn in safety science.
- **Sterman (2006), "Learning from Evidence in a Complex World".** Why learning from experience fails in complex systems, and what simulation can add to evidence-based practice.

The evidence track

These are the studies on which the book's grades most depend. Read them to judge the grades for yourself.

- **Sweeney and Sterman (2000)** and **Cronin, Gonzalez and Sterman (2009)**: the bathtub experiments and the systematic search for why well-educated adults fail them.
- **Fischer, Degen and Funke (2015)**: the most important boundary condition on that finding, showing how much depends on presenting problems as graphs.
- **Sterman (1989a)** and **Croson and Donohue (2006)**: the beer game, and the demonstration that the bullwhip survives when its operational causes are removed.
- **Moxnes (1998)**: fisheries professionals overinvesting in a stock they owned outright.
- **Stadler et al. (2015)**: the meta-analysis of complex problem solving and intelligence.
- **Scott, Cavana and Cameron (2016)**: what group model building has and has not been shown to deliver.

- **Allender et al. (2021)**: the WHO STOPS trial, the strongest test of a community systems intervention.
- **Goncalves Filho, Jun and Waterson (2019)**: four analyses of one accident, and how little two methods agreed.

The lineage track

For readers who want the primary sources of the ideas traced in Part II, in roughly the order in which they appeared.

- **Rosenblueth, Wiener and Bigelow (1943)**: purpose defined through feedback, the founding paper of cybernetics.
- **Wiener (1948),** ***Cybernetics***, and **Ashby (1956),** ***An Introduction to Cybernetics***: the two founding books, the second containing the law of requisite variety.
- **Weaver (1948), "Science and complexity"**: the distinction between disorganised and organised complexity.
- **von Bertalanffy (1968),** ***General System Theory***: the search for laws common to many disciplines.
- **Simon (1962), "The architecture of complexity"**: hierarchy, near-decomposability and the parable of the two watchmakers.
- **Forrester (1961),** ***Industrial Dynamics***: the founding text of system dynamics.
- **Churchman (1968),** ***The Systems Approach***, and **Checkland (1981),** ***Systems Thinking, Systems Practice***: the turn toward systems as ways of seeing, and the birth of the soft tradition.
- **Beer (1972),** ***Brain of the Firm***: the first full statement of the Viable System Model.
- **Heims (1991),** ***The Cybernetics Group***: the history of the conferences where cybernetics took shape.

The critics' track

Every tradition is best understood through its strongest critics. These are the ones this book found most useful.

- **Nordhaus (1973)** and **Freeman, Jahoda et al. (1973), *Models of Doom***: the economists' and the Sussex group's cases against the world models.
- **Kadanoff (1971)** and **Lee (1973), "Requiem for Large-Scale Models"**: the critique of Urban Dynamics and of large planning models generally.
- **Oreskes, Shrader-Frechette and Belitz (1994)**: why models of open systems cannot be validated, only evaluated.
- **Scott (1998), *Seeing Like a State***: how schemes to order society from above erase the practical knowledge that makes it work.
- **Grieves (2008)**: the case for abandoning the learning organisation.
- **Richardson (1986), "Problems with causal-loop diagrams"**: a warning from inside the field that every diagrammer should read.
- **Boettiger and Hastings (2012)**: the limits of early-warning signals.
- **Ackermann (2012)**: problem structuring methods put in the dock, with a frank look at their thin evaluation record.

Appendix D. Methods Toolkit

These cards summarise the practitioner tools that close Chapters 1 to 19. They are reminders, not substitutes: each chapter explains its tool, works an example and gives the reasoning behind the grade. Grades refer to what the evidence says about the problem a tool addresses or the benefit it can deliver, as set out in Appendix A. Chapter 20 has no card of its own: its practitioner's code draws on these nineteen tools, and it is set out, with its one-page version, in that chapter.

1. Stock-flow sketch and accumulation check (Chapters 1 and 2)

Reach for this whenever someone reports a trend. Draw the stock as a box, one arrow in and one out, with units. Then ask what the stock is, what the flows are and what the net flow is doing now. The classic slip is reading the level off the latest change in the inflow. The error it guards against is grade A.

2. Format rules for dynamic data (Chapter 2)

For anyone preparing charts for a decision. Plot accumulations explicitly, set every flow chart beside the stock chart it feeds, and restate the problem in plain words. A clean graph is not self-explanatory. The format effects are laboratory findings (B to C); their size in workplace displays is less certain.

3. Supply-line accounting and the wait-for-the-delay rule (Chapter 3)

Before ordering, hiring or correcting again, total what is already in the pipeline and estimate how long it takes to arrive. Act on the gap that will remain, not the gap visible today. The misperception is grade A; the remedy's benefit outside the laboratory is C.

4. Dörner pre-mortem checklist (Chapter 4)

Run it before launching a complex intervention. Imagine the plan has failed, then look for neglected side-effects, actions fired off without follow-up, drift from topic to topic and retreat into small solvable tasks. It is a heuristic built from an unreplicated taxonomy, so treat it as C.

5. Learning-environment design brief (Chapter 5)

Use when experience is not teaching people what they need. Decide whether a practice simulator is warranted, then make sure it compresses delays, permits repeated trials and gives feedback about structure as well as results. Unstructured practice improves scores without insight. The evidence grade is C.

6. Requisite-variety audit (Chapter 6)

List the disturbances a control arrangement faces and the distinct responses available to it, and compare the two lists. Formal procedures that nobody can actually invoke do not count as responses. This is theory, with no outcome studies behind it.

7. Boundary-and-hierarchy sketch (Chapter 7)

Mark where the system boundary falls and why, then identify subsystems whose internal ties are much stronger than their ties to each other; those can often be handled separately. This too is theory, but it is grounded in Simon's argument about near-decomposability.

8. System dynamics modelling steps (Chapter 8)

Articulate the problem, state a dynamic hypothesis, formulate a model, test it, and only then design policy. The costly shortcut is leaping from hypothesis to recommendation. The sequence is the field's practice standard rather than an evidence-graded method.

9. CATWOE and boundary-critique questions (Chapter 9)

Draft a root definition with its CATWOE elements, then ask Ulrich's boundary questions twice, once describing how things are and once how they ought to be. Watch for the powerful framing that nobody questions. C for use, D for comparative effectiveness.

10. Causal loop diagram protocol with stock check (Chapter 10)

Sketch the behaviour over time first, name variables as quantities with units and write out each polarity sentence in full. Box every stock, draw one stock-flow fragment with numbers, and record each apparent lever as a hypothesis. The protocol rests on theory and expert guidance and is untested as a procedure.

11. Two-workshop group model building with measures (Chapter 11)

Collect individual diagrams, action lists and a stock-flow item before the first workshop and after the second, add a consensus questionnaire, and check agreed actions and predictions at six and twelve months. B for the proximal outcomes it captures; downstream effects remain unmeasured.

12. Leverage question set and lever-claim audit (Chapter 12)

Walk the twelve places from bottom to top to generate options, then ignore the ranking and judge each option on evidence, delay, likely resistance and side-effects. For any claimed lever, ask where its effect size came from and what it was compared with. The ranking is graded D, but the list remains useful as a prompt.

13. Archetypes as testable hypotheses (Chapter 13)

Translate a recognised archetype into predicted time series, name its delay and hidden stocks, collect data covering the delay, and fix the disconfirming pattern in advance. Pair it with a quick psychological-safety check. Archetype diagnosis is D; psychological safety itself is B.

14. Assessment blueprint for systems-thinking skill (Chapter 14)

Combine a performance task, a published self-report scale, a transfer task in another format and a short reasoning test, measured before, after and months later, ideally with a comparison group. Read divergence between the scores as information. The evidence grade is C.

15. Sceptic's guide questions for any model (Chapter 15)

When shown a model, ask what its boundary leaves out, which assumptions drive the result, how it was tested and what purpose it was built for (Sterman 1991). A model that fits past data can still have the wrong structure. The questions rest on theory rather than tested outcomes.

16. Viable System Model diagnostic (Chapter 16)

Take a team you know and locate each of Systems 1 to 5 in it, noting any that are missing, duplicated or starved of information. Useful for provoking questions about coordination and autonomy; there is no outcome evidence, so D.

17. Evaluation plan for systems interventions (Chapter 17)

Pre-specify outcomes, choose a stepped-wedge or other comparison design, and measure how systemic the intervention actually was in practice. The plan follows current MRC and NIHR guidance (Skivington et al. 2021; Egan et al. 2019), which is consensus guidance rather than tested method.

18. Commons governance audit (Chapter 18)

Check a shared-resource institution against Ostrom's eight design principles, then run the misperception check: do users grasp the stock's regeneration and the delays in its response? B for common-pool resources; D if stretched to organisations in general.

19. Multi-method incident review (Chapter 19)

Analyse the same incident with two different methods, using separate analysts where possible, and report where their findings overlap and where they diverge. Low overlap says something about the methods themselves and belongs in the report. The evidence grade is C.

Appendix E. A Note on Sources and Verification

A book that grades other people's evidence should say how its own was checked. Every work cited here appears in a working bibliography of 344 sources, and each entry was checked against a live record before it was used. Journal articles and chapters with a DOI, 255 of them, were matched against their Crossref metadata, with abstracts read from OpenAlex, Crossref or Semantic Scholar where available. Books were checked against Open Library records for author, title, first publication year and publisher. Reports and web documents were checked against the issuing pages themselves, including GOV.UK, WHO IRIS, PubMed and publisher sites. Where a record disagreed with the conventional citation, as when a journal's online year differs from its print year, the bibliography records the difference.

A second pass on 27 September 2026 went back to the sources themselves for figures and quotations that the first pass had taken from abstracts or secondary accounts. Its results changed the book in three ways. It confirmed exact wording for a small set of quotations, and only those are quoted directly; everything else is paraphrased. It resolved 22 of 24 items that had been flagged as unverified. And it corrected one error of substance: earlier project notes had said that Turner (2008) denied his comparison of World3 with observed data amounted to validation, when the working-paper version says it "provides a degree of validation of the World3 model". Chapter 15 uses the corrected reading.

Two works remain unverified and are not cited anywhere: a 1980 paper by Forrester and Senge on tests for building confidence in system dynamics models, and a 1982 paper by Jackson on the nature of soft systems thinking. Both are known only from search indexes, and no record of either could be retrieved.

Several figures carry conditions that the text keeps. The Foresight obesity figures are extrapolations of past trends, and the book always says that half of UK adults "could be" obese by 2050, never that they will be. The Turner quotations are from the CSIRO working-paper version of his study. Any figures from the Diehl and Sterman experiment are given as working-paper figures, because the published article could not be accessed. Some studies are cited for their design or topic only, with no results stated, because their results could not be read in the primary source. And the percentages attached to Meadows's electric-meter and pollution-inventory examples are reported as her own unsourced claims, since the essay gives no source for either.

Some limits on the checking should also be recorded. Several services were rate-limited or blocked, publisher full texts were sometimes reachable only through author or repository copies, and no attempt was made to bypass access controls. Where a detail could not be confirmed, it was left out.

A final pass during editing checked the remaining descriptive claims against full texts where they could be reached: Simon's 1962 paper for the parable of the two watchmakers, the Cox, Arnold and Villamayor Tomás review for the coding of Ostrom's design principles, the text of Meadows's essay for her examples and her closing caution, and Beer's own account of the Viable System Model. It also checked the print year of every article whose online and print years differ, and the in-text citations now give the print year throughout. Where the soft systems literature could be checked only through secondary accounts, as for the CATWOE mnemonic, the text follows the most careful of those accounts.

References

This list contains every work cited in the text, in author-date form. Where a journal article appeared online before its print issue, the print year is used in the text and the online year is noted.

Abdel-Hamid, T., Ankel, F., Battle-Fisher, M., Gibson, B., Gonzalez-Parra, G., Jalali, M. et al. (2014). Public and health professionals' misconceptions about the dynamics of body weight gain/loss. *System Dynamics Review* 30(1-2), 58-74.

Abson, D. J., Fischer, J., Leventon, J., Newig, J., Schomerus, T., Vilsmaier, U. et al. (2017). Leverage points for sustainability transformation. *Ambio* 46(1), 30-39. First published online 2016.

Ackermann, F. (2012). Problem structuring methods 'in the Dock': Arguing the case for Soft OR. *European Journal of Operational Research* 219(3), 652-658.

Ackoff, R. L. (1971). Towards a System of Systems Concepts. *Management Science* 17(11), 661-671.

Ackoff, R. L. (1979). The Future of Operational Research is Past. *Journal of the Operational Research Society* 30(2), 93-104.

Adam, T. & de Savigny, D. (2012). Systems thinking for strengthening health systems in LMICs: need for a paradigm shift. *Health Policy and Planning* 27(suppl 4), iv1-iv3.

Allender, S., Millar, L., Hovmand, P., Bell, C., Moodie, M., Carter, R. et al. (2016). Whole of Systems Trial of Prevention Strategies for Childhood Obesity: WHO STOPS Childhood Obesity. *International Journal of Environmental Research and Public Health* 13(11), 1143.

Allender, S., Orellana, L., Crooks, N., Bolton, K. A., Fraser, P., Brown, A. D. et al. (2021). Four-Year Behavioral, Health-Related Quality of Life, and BMI Outcomes from a Cluster Randomized Whole of Systems Trial of Prevention Strategies for Childhood Obesity. *Obesity* 29(6), 1022-1035.

Allender, S., Owen, B., Kuhlberg, J., Lowe, J., Nagorcka-Smith, P., Whelan, J. et al. (2015). A Community Based Systems Diagram of Obesity Causes. *PLOS ONE* 10(7), e0129683.

Andersen, D. F. & Richardson, G. P. (1997). Scripts for group model building. *System Dynamics Review* 13(2), 107-129.

Andersen, D. F., Vennix, J. A. M., Richardson, G. P. & Rouwette, E. A. J. A. (2007). Group model building: problem structuring, policy simulation and decision support. *Journal of the Operational Research Society* 58(5), 691-694.

Anderson, P. W. (1972). More Is Different. *Science* 177(4047), 393-396.

Arango Aramburo, S., Castañeda Acevedo, J. A. & Olaya Morales, Y. (2012). Laboratory experiments in the system dynamics field. *System Dynamics Review* 28(1), 94-106.

Argyris, C. (1977). Double loop learning in organizations. *Harvard Business Review*, September 1977.

Argyris, C. & Schön, D.A. (1978). Organizational Learning: A Theory of Action Perspective. Reading, MA: Addison-Wesley.

Armstrong McKay, D. I., Staal, A., Abrams, J. F., Winkelmann, R., Sakschewski, B., Loriani, S. et al. (2022). Exceeding 1.5°C global warming could trigger multiple climate tipping points. *Science* 377(6611).

Arnold, R. D. & Wade, J. P. (2015). A Definition of Systems Thinking: A Systems Approach. *Procedia Computer Science* 44, 669-678.

Arthur, W. B. (1989). Competing Technologies, Increasing Returns, and Lock-In by Historical Events. *The Economic Journal* 99(394), 116.

Ashby, W.R. (1952). Design for a Brain. London: Chapman & Hall.

Ashby, W.R. (1956). An Introduction to Cybernetics. London: Chapman & Hall.

Ashby, W. R. (1958). Requisite variety and its implications for the control of complex systems. *Cybernetica* 1(2), 83-99. Reprinted in *Facets of Systems Science* (1991), pp. 405-417.

Assaraf, O. B. & Orion, N. (2005). Development of system thinking skills in the context of earth system education. *Journal of Research in Science Teaching* 42(5), 518-560.

Atkinson, J. A., Wells, R., Page, A., Dominello, A., Haines, M. & Wilson, A. (2015). Applications of system dynamics modelling to support health policy. *Public Health Research and Practice* 25(3).

Baggio, J. A., Barnett, A. J., Perez-Ibarra, I., Brady, U., Ratajczyk, E., Rollins, N. et al. (2016). Explaining success and failure in the commons: the configural nature of Ostrom's institutional design principles. *International Journal of the Commons* 10(2), 417.

Bagnall, A. M., Radley, D., Jones, R., Gately, P., Nobles, J., Van Dijk, M. et al. (2019). Whole systems approaches to obesity and other complex public health challenges: a systematic review. *BMC Public Health* 19(1).

Bardi, U. (2011). The Limits to Growth Revisited. Springer New York (SpringerBriefs in Energy).

Barlas, Y. (1996). Formal aspects of model validity and validation in system dynamics. *System Dynamics Review* 12(3), 183-210.

Bateson, G. (1972). Steps to an Ecology of Mind. San Francisco: Chandler.

Beer, S. (1972). Brain of the Firm. London: Allen Lane / New York: Herder and Herder.

Beer, S. (1975). Platform for Change. London: Wiley.

Beer, S. (1979). The Heart of Enterprise. Chichester: Wiley.

Beer, S. (1985). Diagnosing the System for Organizations. Chichester: Wiley.

Boettiger, C. & Hastings, A. (2012). Quantifying limits to detection of early warning for critical transitions. *Journal of The Royal Society Interface* 9(75), 2527-2539.

Boulding, K. E. (1956). General Systems Theory—The Skeleton of Science. *Management Science* 2(3), 197-208.

Braithwaite, J., Wears, R. L. & Hollnagel, E. (2015). Resilient health care: turning patient safety on its head. *International Journal for Quality in Health Care* 27(5), 418-420.

Brand, F. S. & Jax, K. (2007). Focusing the Meaning(s) of Resilience: Resilience as a Descriptive Concept and a Boundary Object. *Ecology and Society* 12(1).

Brehmer, B. (1992). Dynamic decision making: Human control of complex systems. *Acta Psychologica* 81(3), 211-241.

Brehmer, B. & Dörner, D. (1993). Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Computers in Human Behavior* 9(2-3), 171-184.

Brunstein, A., Gonzalez, C. & Kanter, S. (2010). Effects of domain experience in the stock–flow failure. *System Dynamics Review* 26(4), 347-354.

Butland, B., Jebb, S., Kopelman, P., McPherson, K., Thomas, S., Mardell, J. & Parry, V. (2007). Tackling Obesities: Future Choices, Project Report (2nd ed.). London: Government Office for Science (Foresight).

Cabrera, D. & Cabrera, L. (2015). Systems Thinking Made Simple: New Hope for Solving Wicked Problems. Ithaca: Odyssean.

Cabrera, D., Colosi, L. & Lobdell, C. (2008). Systems thinking. *Evaluation and Program Planning* 31(3), 299-310.

Caldwell, R. (2012). Systems Thinking, Organizational Change and Agency: A Practice Theory Critique of Senge's Learning Organization. *Journal of Change Management* 12(2), 145-164.

Camelia, F. & Ferris, T. L. J. (2018). Validation Studies of a Questionnaire Developed to Measure Students' Engagement With Systems Thinking. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 48(4), 574-585.

Capelo, C. & Dias, J. F. (2009). A system dynamics-based simulation experiment for testing mental model and performance effects of using the balanced scorecard. *System Dynamics Review* 25(1), 1-34.

Carey, G., Malbon, E., Carey, N., Joyce, A., Crammond, B. & Carey, A. (2015). Systems science and systems thinking for public health: a systematic review of the field. *BMJ Open* 5(12), e009002.

Castro, R. (2012). Arguments on the Imminence of Global Collapse Are Premature when Based on Simulation Models. *GAIA - Ecological Perspectives for Science and Society* 21(4), 271-273.

Chan, K. M. A., Boyd, D. R., Gould, R. K., Jetzkowitz, J., Liu, J., Muraca, B. et al. (2020). Levers and leverage points for pathways to sustainability. *People and Nature* 2(3), 693-717.

Checkland, P. (1981). Systems Thinking, Systems Practice. Chichester: Wiley.

Checkland, P. (2000). Soft systems methodology: a thirty year retrospective. *Systems Research and Behavioral Science* 17(S1), S11-S58.

Checkland, P. & Poulter, J. (2006). Learning for Action: A Short Definitive Account of Soft Systems Methodology. Chichester: Wiley.

Checkland, P. & Scholes, J. (1990). Soft Systems Methodology in Action. Chichester: Wiley.

Chi, M. T. H., Roscoe, R. D., Slotta, J. D., Roy, M. & Chase, C. C. (2012). Misconceived Causal Explanations for Emergent Processes. *Cognitive Science* 36(1), 1-61. First published online 2011.

Chughtai, S. & Blanchet, K. (2017). Systems thinking in public health: a bibliographic contribution to a meta-narrative review. *Health Policy and Planning*, czw159.

Churchman, C.W. (1968). The Systems Approach. New York: Dell.

Conant, R. C. & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science* 1(2), 89-97.

Cook, R.I. (n.d.). How Complex Systems Fail. Cognitive Technologies Laboratory, University of Chicago.

Cox, M., Arnold, G. & Villamayor Tomás, S. (2010). A Review of Design Principles for Community-based Natural Resource Management. *Ecology and Society* 15(4).

Cronin, M. A. & Gonzalez, C. (2007). Understanding the building blocks of dynamic systems. *System Dynamics Review* 23(1), 1-17.

Cronin, M. A., Gonzalez, C. & Sterman, J. D. (2009). Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. *Organizational Behavior and Human Decision Processes* 108(1), 116-130.

Croson, R. & Donohue, K. (2006). Behavioral Causes of the Bullwhip Effect and the Observed Value of Inventory Information. *Management Science* 52(3), 323-336.

Croson, R., Donohue, K., Katok, E. & Sterman, J. (2014). Order Stability in Supply Chains: Coordination Risk and the Role of Coordination Stock. *Production and Operations Management* 23(2), 176-196.

Currie, C. S., Fowler, J. W., Kotiadis, K., Monks, T., Onggo, B. S., Robertson, D. A. et al. (2020). How simulation modelling can help reduce the impact of COVID-19. *Journal of Simulation* 14(2), 83-97.

Darabi, N. & Hosseinichimeh, N. (2020). System dynamics modeling in health and medicine: a systematic literature review. *System Dynamics Review* 36(1), 29-73.

Davis, A. C. & Stroink, M. L. (2016). The Relationship between Systems Thinking and the New Ecological Paradigm. *Systems Research and Behavioral Science* 33(4), 575-586. First published online 2015.

Dekker, S. (2010). Drift into Failure: From Hunting Broken Components to Understanding Complex Systems. Farnham: Ashgate.

Dekker, S., Cilliers, P. & Hofmeyr, J. H. (2011). The complexity of failure: Implications of complexity theory for safety investigations. *Safety Science* 49(6), 939-945.

Deming, W.E. (1982). Out of the Crisis. Cambridge, MA: MIT Center for Advanced Engineering Study.

de Savigny, D. & Adam, T. (eds.) (2009). Systems Thinking for Health Systems Strengthening. Geneva: Alliance for Health Policy and Systems Research / WHO.

Diehl, E. & Sterman, J. D. (1995). Effects of Feedback Complexity on Dynamic Decision Making. *Organizational Behavior and Human Decision Processes* 62(2), 198-215.

Dietz, T., Ostrom, E. & Stern, P. C. (2003). The Struggle to Govern the Commons. *Science* 302(5652), 1907-1912.

Diez Roux, A. V. (2011). Complex Systems Thinking and Current Impasses in Health Disparities Research. *American Journal of Public Health* 101(9), 1627-1634.

Dolansky, M. A., Moore, S. M., Palmieri, P. A. & Singh, M. K. (2020). Development and Validation of the Systems Thinking Scale. *Journal of General Internal Medicine* 35(8), 2314-2320.

Dörner, D. (1980). On the Difficulties People Have in Dealing With Complexity. *Simulation & Games* 11(1), 87-106.

Dörner, D. (1996). The Logic of Failure: Why Things Go Wrong and What We Can Do to Make Them Right (R. & R. Kimber, trans.). New York: Metropolitan Books.

Dörner, D. & Funke, J. (2017). Complex Problem Solving: What It Is and What It Is Not. *Frontiers in Psychology* 8.

Dörner, D., Kreuzig, H.W., Reither, F., Stäudel, T. et al. (eds.) (1983). Lohhausen: Vom Umgang mit Unbestimmtheit und Komplexität. Bern: Hans Huber.

Dorninger, C., Abson, D. J., Apetrei, C. I., Derwort, P., Ives, C. D., Klaniecki, K. et al. (2020). Leverage points for sustainability transformation: a review on interventions in food and energy systems. *Ecological Economics* 171, 106570.

Doyle, J. K. & Ford, D. N. (1998). Mental models concepts for system dynamics research. *System Dynamics Review* 14(1), 3-29.

Dugan, K. E., Mosyjowski, E. A., Daly, S. R. & Lattuca, L. R. (2022). Systems thinking assessments in engineering: A systematic literature review. *Systems Research and Behavioral Science* 39(4), 840-866. First published online 2021.

Dutt, V. & Gonzalez, C. (2012). Decisions from experience reduce misconceptions about climate change. *Journal of Environmental Psychology* 32(1), 19-29.

Edmondson, A. (1999). Psychological Safety and Learning Behavior in Work Teams. *Administrative Science Quarterly* 44(2), 350-383.

Egan, M., McGill, E., Penney, T., Anderson de Cuevas, R., Er, V., Orton, L., White, M., Lock, K., Cummins, S., Savona, N., Whitehead, M., Popay, J., Smith, R., Meier, P., De Vocht, F., Marks, D., Andreeva, M., Rutter, H. & Petticrew, M. (2019). NIHR SPHR Guidance on Systems Approaches to Local Public Health Evaluation. Part 2: What to consider when planning a systems evaluation. London: NIHR School for Public Health Research. Published 25 March 2019.

Ellinger, A. D., Ellinger, A. E., Yang, B. & Howton, S. W. (2002). The relationship between the learning organization concept and firms' financial performance: An empirical assessment. *Human Resource Development Quarterly* 13(1), 5-22.

Espejo, R. & Harnden, R. (eds.) (1989). The Viable System Model: Interpretations and Applications of Stafford Beer's VSM.

Espinosa, A., Harnden, R. & Walker, J. (2008). A complexity approach to sustainability – Stafford Beer revisited. *European Journal of Operational Research* 187(2), 636-651.

Evagorou, M., Korfiatis, K., Nicolaou, C. & Constantinou, C. (2009). An Investigation of the Potential of Interactive Simulations for Developing System Thinking Skills in Elementary School: A case study with fifth-graders and sixth-graders. *International Journal of Science Education* 31(5), 655-674.

Featherston, C. R. & Doolan, M. (2012). A critical review of the criticisms of system dynamics. *Proceedings of the 30th International Conference of the System Dynamics Society*, St. Gallen.

Finegood, D. T., Merth, T. D. & Rutter, H. (2010). Implications of the Foresight Obesity System Map for Solutions to Childhood Obesity. *Obesity* 18(S1).

Fischer, H., Degen, C. & Funke, J. (2015). Improving Stock-Flow Reasoning With Verbal Formats. *Simulation & Gaming* 46(3-4), 255-269.

Fischer, H. & Gonzalez, C. (2016). Making Sense of Dynamic Systems: How Our Understanding of Stocks and Flows Depends on a Global Perspective. *Cognitive Science* 40(2), 496-512. First published online 2015.

Fischer, J. & Riechers, M. (2019). A leverage points perspective on sustainability. *People and Nature* 1(1), 115-120.

Flood, R.L. (1999). Rethinking the Fifth Discipline: Learning within the Unknowable. London: Routledge.

Flood, R. L. & Jackson, M. C. (1991). Creative Problem Solving: Total Systems Intervention. Chichester: Wiley.

Folke, C. (2006). Resilience: The emergence of a perspective for social–ecological systems analyses. *Global Environmental Change* 16(3), 253-267.

Forrester, J. W. (1958). Industrial dynamics: a major breakthrough for decision makers. *Harvard Business Review*, July-August 1958. Reprinted in *The Roots of Logistics* (Springer, 2012), pp. 141-172.

Forrester, J.W. (1961). Industrial Dynamics. Cambridge, MA: MIT Press.

Forrester, J.W. (1969). Urban Dynamics. Cambridge, MA: MIT Press.

Forrester, J. W. (1971a). Counterintuitive behavior of social systems. *Theory and Decision* 2(2), 109-140.

Forrester, J. W. (1971b). World Dynamics. Cambridge, MA: Wright-Allen Press.

Forrester, J. W. (1987). Lessons from system dynamics modeling. *System Dynamics Review* 3(2), 136-149.

Forrester, J. W. (1993). System dynamics as an organizing framework for pre-college education. *System Dynamics Review* 9(2), 183-194.

Forrester, J. W. (2007). System dynamics—a personal view of the first fifty years. *System Dynamics Review* 23(2-3), 345-358.

Forrester, J. W., Low, G. W. & Mass, N. J. (1974). The debate on World Dynamics: A response to Nordhaus. *Policy Sciences* 5(2), 169-190.

Frazier, M. L., Fainshmidt, S., Klinger, R. L., Pezeshkan, A. & Vracheva, V. (2017). Psychological Safety: A Meta-Analytic Review and Extension. *Personnel Psychology* 70(1), 113-165. First published online 2016.

Freeman, C., Jahoda, M. et al. (eds.) (1973). Models of Doom: A Critique of The Limits to Growth. New York: Universe Books.

Funke, J. (1993). Chapter 14 Microworlds Based on Linear Equation Systems: A New Approach to Complex Problem Solving and Experimental Results. *Advances in Psychology*, 313-330.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning* 7(1), 69-89.

Funke, J. (2010). Complex problem solving: a case for complex cognition? *Cognitive Processing* 11(2), 133-142. First published online 2009.

Galea, S., Riddle, M. & Kaplan, G. A. (2010). Causal thinking and complex system approaches in epidemiology. *International Journal of Epidemiology* 39(1), 97-106. First published online 2009.

Garvin, D. A. (1993). Building a learning organization. *Harvard Business Review* 71(4), 78-91.

Garvin, D. A., Edmondson, A. C. & Gino, F. (2008). Is yours a learning organization? *Harvard Business Review* 86(3), 109-116, 134.

Gerritsen, S., Harré, S., Rees, D., Renker-Darby, A., Bartos, A. E., Waterlander, W. E. et al. (2020). Community Group Model Building as a Method for Engaging Participants and Mobilising Action in Public Health. *International Journal of Environmental Research and Public Health* 17(10), 3457.

Ghaffarzadegan, N., Lyneis, J. & Richardson, G. P. (2011). How small system dynamics models can help the public policy process. *System Dynamics Review* 27(1), 22-44. First published online 2010.

Gilissen, M. G. R., Knippels, M. C. P. J. & van Joolingen, W. R. (2020). Bringing systems thinking into the classroom. *International Journal of Science Education* 42(8), 1253-1280.

Goh, S. C., Elliott, C. & Quon, T. K. (2012). The relationship between learning capability and organizational performance. *The Learning Organization* 19(2), 92-108.

Goncalves Filho, A. P., Jun, G. T. & Waterson, P. (2019). Four studies, two methods, one accident – An examination of the reliability and validity of Accimap and STAMP for accident analysis. *Safety Science* 113, 310-317.

Gonzalez, C., Vanyukov, P. & Martin, M. K. (2005). The use of microworlds to study dynamic decision making. *Computers in Human Behavior* 21(2), 273-286.

Gonzalez, C. & Wong, H. (2011). Understanding stocks and flows through analogy. *System Dynamics Review* 28(1), 3-27.

Gray, J. N., Pessel, D. & Varaiya, P. P. (1972). A Critique of Forrester's Model of an Urban Area. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-2(2), 139-144.

Greenhalgh, T. & Papoutsi, C. (2018). Studying complexity in health services research: desperately seeking an overdue paradigm shift. *BMC Medicine* 16(1).

Greiff, S., Wüstenberg, S. & Funke, J. (2012). Dynamic Problem Solving. *Applied Psychological Measurement* 36(3), 189-213.

Grieves, J. (2008). Why we should abandon the idea of the learning organization. *The Learning Organization* 15(6), 463-473.

Grohs, J. R., Kirk, G. R., Soledad, M. M. & Knight, D. B. (2018). Assessing systems thinking: A tool to measure complex reasoning through ill-structured problems. *Thinking Skills and Creativity* 28, 110-130.

Güss, C. D. & Dörner, D. (2011). Cultural differences in dynamic decision-making strategies in a non-linear, time-delayed task. *Cognitive Systems Research* 12(3-4), 365-376.

Güss, C. D., Tuason, M. T. & Gerhard, C. (2010). Cross-National Comparisons of Complex Problem-Solving Strategies in Two Microworlds. *Cognitive Science* 34(3), 489-520.

Guy, S., Kashima, Y., Walker, I. & O'Neill, S. (2013). Comparing the atmosphere to a bathtub: effectiveness of analogy for reasoning about accumulation. *Climatic Change* 121(4), 579-594.

Hall, K. D., Sacks, G., Chandramohan, D., Chow, C. C., Wang, Y. C., Gortmaker, S. L. et al. (2011). Quantification of the effect of energy imbalance on bodyweight. *The Lancet* 378(9793), 826-837.

Hammond, D. (2003). The Science of Synthesis: Exploring the Social Implications of General Systems Theory. Boulder: University Press of Colorado.

Hardin, G. (1968). The Tragedy of the Commons. *Science* 162(3859), 1243-1248.

Hawe, P., Shiell, A. & Riley, T. (2009). Theorising Interventions as Events in Systems. *American Journal of Community Psychology* 43(3-4), 267-276.

Heims, S.J. (1991). The Cybernetics Group. Cambridge, MA: MIT Press.

Herrington, G. (2021). Update to limits to growth: Comparing the World3 model with empirical data. *Journal of Industrial Ecology* 25(3), 614-626. First published online 2020.

Hirsch, G., Homer, J., Evans, E. & Zielinski, A. (2010). A System Dynamics Model for Planning Cardiovascular Disease Interventions. *American Journal of Public Health* 100(4), 616-622.

Hmelo-Silver, C. E., Jordan, R., Eberbach, C. & Sinha, S. (2017). Systems learning with a conceptual representation: a quasi-experimental study. *Instructional Science* 45(1), 53-72. First published online 2016.

Hmelo-Silver, C. E. & Pfeffer, M. G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cognitive Science* 28(1), 127-138.

Holland, J.H. (1995). Hidden Order: How Adaptation Builds Complexity. Reading, MA: Addison-Wesley.

Holling, C. & Meffe, G. K. (1996). Command and Control and the Pathology of Natural Resource Management. *Conservation Biology* 10(2), 328-337.

Holling, C. S. (1973). Resilience and Stability of Ecological Systems. *Annual Review of Ecology and Systematics* 4(1), 1-23.

Hollnagel, E. (2012). FRAM: The Functional Resonance Analysis Method. Farnham: Ashgate / CRC.

Hollnagel, E. (2014). Safety-I and Safety-II: The Past and Future of Safety Management. Farnham: Ashgate / CRC.

Hollnagel, E., Woods, D.D. & Leveson, N. (eds.) (2006). Resilience Engineering: Concepts and Precepts. Aldershot: Ashgate.

Homer, J. B. & Hirsch, G. B. (2006). System Dynamics Modeling for Public Health: Background and Opportunities. *American Journal of Public Health* 96(3), 452-458.

Hopper, M. & Stave, K. A. (2008). Assessing the effectiveness of systems thinking interventions in the classroom. *Proceedings of the 26th International Conference of the System Dynamics Society*, Athens.

Hovmand, P. S. (2014). Community Based System Dynamics. Springer New York.

Hovmand, P. S., Andersen, D. F., Rouwette, E., Richardson, G. P., Rux, K. & Calhoun, A. (2012). Group Model-Building 'Scripts' as a Collaborative Planning Tool. *Systems Research and Behavioral Science* 29(2), 179-193.

Hulme, A., Stanton, N. A., Walker, G. H., Waterson, P. & Salmon, P. M. (2019). What do applications of systems thinking accident analysis methods tell us about accident causation? A systematic review of applications between 1990 and 2018. *Safety Science* 117, 164-183.

Iflaifel, M., Lim, R. H., Ryan, K. & Crowley, C. (2020). Resilient Health Care: a systematic review of conceptualisations, study methods and factors that develop resilience. *BMC Health Services Research* 20(1).

Jackson, M. C. (1988). An appreciation of stafford beer's 'viable system' viewpoint on managerial practice. *Journal of Management Studies* 25(6), 557-573.

Jackson, M.C. (2003). Systems Thinking: Creative Holism for Managers. Chichester: Wiley.

Jackson, M.C. (2019). Critical Systems Thinking and the Management of Complexity. Hoboken: Wiley.

Jacobson, M. J. (2001). Problem solving, cognition, and complex systems: Differences between experts and novices. *Complexity* 6(3), 41-49.

Janssen, M. A., Holahan, R., Lee, A. & Ostrom, E. (2010). Lab Experiments for the Study of Social-Ecological Systems. *Science* 328(5978), 613-617.

Johnston, L. M., Matteson, C. L. & Finegood, D. T. (2014). Systems Science and Obesity Policy: A Novel Framework for Analyzing and Rethinking Population-Level Planning. *American Journal of Public Health* 104(7), 1270-1278.

Jones, A. P., Homer, J. B., Murphy, D. L., Essien, J. D. K., Milstein, B. & Seville, D. A. (2006). Understanding Diabetes Population Dynamics Through Simulation Modeling and Experimentation. *American Journal of Public Health* 96(3), 488-494.

Kadanoff, L. P. (1971). From simulation model to public policy. *SIMULATION* 16(6), 261-268.

Kainz, D. & Ossimitz, G. (2002). Can students learn stock-flow-thinking? An empirical investigation. *Proceedings of the 20th International Conference of the System Dynamics Society*, Palermo.

Kampmann, C. E. & Sterman, J. D. (2014). Do markets mitigate misperceptions of feedback? *System Dynamics Review* 30(3), 123-160.

Kim, D. H. (1993). The link between individual and organizational learning. *Sloan Management Review*, Fall 1993.

Kim, K., Watkins, K. E. & Lu, Z. (. (2017). The impact of a learning organization on performance. *European Journal of Training and Development* 41(2), 177-193.

Kline, R.R. (2015). The Cybernetics Moment: Or Why We Call Our Age the Information Age. Baltimore: Johns Hopkins University Press.

Kluge, A. (2008). Performance Assessments With Microworlds and Their Difficulty. *Applied Psychological Measurement* 32(2), 156-180.

Knai, C., Petticrew, M., Mays, N., Capewell, S., Cassidy, R., Cummins, S. et al. (2018). Systems Thinking as a Framework for Analyzing Commercial Determinants of Health. *The Milbank Quarterly* 96(3), 472-498.

Kofman, F. & Senge, P. M. (1993). Communities of commitment: The heart of learning organizations. *Organizational Dynamics* 22(2), 5-23.

Kopainsky, B. & Sawicka, A. (2011). Simulator-supported descriptions of complex dynamic problems: experimental results on task performance and system understanding. *System Dynamics Review* 27(2), 142-172. First published online 2010.

Korzilius, H., Raaijmakers, S., Rouwette, E. & Vennix, J. (2014). Thinking aloud while solving a stock-flow task: Surfacing the correlation heuristic and other reasoning patterns. *Systems Research and Behavioral Science* 31(2), 268-279. First published online 2013.

Kretzschmar, A., Neubert, J. C., Wüstenberg, S. & Greiff, S. (2016). Construct validity of complex problem solving: A comprehensive view on different facets of intelligence and school grades. *Intelligence* 54, 55-69.

Kurtz, C. F. & Snowden, D. J. (2003). The new dynamics of strategy: Sense-making in a complex and complicated world. *IBM Systems Journal* 42(3), 462-483.

Lane, D. C. (2000). Should system dynamics be described as a ?hard? or ?deterministic? systems approach?. *Systems Research and Behavioral Science* 17(1), 3-22.

Lane, D. C. (2008). The emergence and use of diagramming in system dynamics: a critical account. *Systems Research and Behavioral Science* 25(1), 3-23.

Lane, D. C., Munro, E. & Husemann, E. (2016). Blending systems thinking approaches for organisational analysis: Reviewing child protection in England. *European Journal of Operational Research* 251(2), 613-623.

Lane, D. C. & Oliva, R. (1998). The greater whole: Towards a synthesis of system dynamics and soft systems methodology. *European Journal of Operational Research* 107(1), 214-235.

Lee, B. Y., Bartsch, S. M., Mui, Y., Haidari, L. A., Spiker, M. L. & Gittelsohn, J. (2017). A systems approach to obesity. *Nutrition Reviews* 75(suppl 1), 94-106.

Lee, D. B. (1973). Requiem for Large-Scale Models. *Journal of the American Institute of Planners* 39(3), 163-178.

Lee, H. L., Padmanabhan, V. & Whang, S. (1997). Information Distortion in a Supply Chain: The Bullwhip Effect. *Management Science* 43(4), 546-558.

Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S. et al. (2008). Tipping elements in the Earth's climate system. *Proceedings of the National Academy of Sciences* 105(6), 1786-1793.

Lenton, T. M., Rockström, J., Gaffney, O., Rahmstorf, S., Richardson, K., Steffen, W. et al. (2019). Climate tipping points — too risky to bet against. *Nature* 575(7784), 592-595.

Leventon, J., Abson, D. J. & Lang, D. J. (2021). Leverage points for sustainability transformations: nine guiding questions for sustainability science and practice. *Sustainability Science* 16(3), 721-726.

Leveson, N. (2004). A new accident model for engineering safer systems. *Safety Science* 42(4), 237-270.

Leveson, N. (2015). A systems approach to risk management through leading safety indicators. *Reliability Engineering & System Safety* 136, 17-34.

Leveson, N. G. (2011). Engineering a Safer World: Systems Thinking Applied to Safety. Cambridge, MA: The MIT Press.

Levin, S. A. (1998). Ecosystems and the Biosphere as Complex Adaptive Systems. *Ecosystems* 1(5), 431-436.

Liu, J., Dietz, T., Carpenter, S. R., Alberti, M., Folke, C., Moran, E. et al. (2007). Complexity of Coupled Human and Natural Systems. *Science* 317(5844), 1513-1516.

Luke, D. A. & Stamatakis, K. A. (2012). Systems Science Methods in Public Health: Dynamics, Networks, and Agents. *Annual Review of Public Health* 33(1), 357-376.

Luna Pinzon, A., Stronks, K., Verhoeff, A., Vaandrager, D., den Hertog, K. & Waterlander, W. (2025). Applying a participatory system dynamics approach to childhood overweight and obesity in the local context: reflections from the LIKE project. *Health Research Policy and Systems* 23(1).

Maani, K. E. & Maharaj, V. (2004). Links between systems thinking and complex decision making. *System Dynamics Review* 20(1), 21-48.

Malhi, L., Karanfil, Ö., Merth, T., Acheson, M., Palmer, A. & Finegood, D. T. (2009). Places to Intervene to Make Complex Food Systems More Healthy, Green, Fair, and Affordable. *Journal of Hunger & Environmental Nutrition* 4(3-4), 466-476.

Marsick, V. J. & Watkins, K. E. (2003). Demonstrating the Value of an Organization's Learning Culture: The Dimensions of the Learning Organization Questionnaire. *Advances in Developing Human Resources* 5(2), 132-151.

Maruyama, M. (1963). The second cybernetics: Deviation-amplifying mutual causal processes. *American Scientist* 51(2), 164-179.

Maturana, H. R. & Varela, F. J. (1980). Autopoiesis and Cognition. Springer Netherlands (Boston Studies in the Philosophy and History of Science).

McGill, E., Marks, D., Er, V., Penney, T., Petticrew, M. & Egan, M. (2020). Qualitative process evaluation from a complex systems perspective: A systematic review and framework for public health evaluators. *PLOS Medicine* 17(11), e1003368.

Meadows, D. (1999). Leverage Points: Places to Intervene in a System. Hartland, VT: Sustainability Institute. Web version: Donella Meadows Project.

Meadows, D.H. (2008). Thinking in Systems: A Primer (D. Wright, ed.). White River Junction, VT: Chelsea Green.

Meadows, D.H., Meadows, D.L., Randers, J. & Behrens, W.W. III (1972). The Limits to Growth. New York: Universe Books.

Meadows, D.H., Randers, J. & Meadows, D.L. (2004). Limits to Growth: The 30-Year Update. White River Junction, VT: Chelsea Green.

Meadows, D.H., Richardson, J. & Bruckmann, G. (1982). Groping in the Dark: The First Decade of Global Modelling. Chichester: Wiley.

Medina, E. (2006). Designing Freedom, Regulating a Nation: Socialist Cybernetics in Allende's Chile. *Journal of Latin American Studies* 38(3), 571-606.

Medina, E. (2011). Cybernetic Revolutionaries: Technology and Politics in Allende's Chile. Cambridge, MA: MIT Press.

Mehren, R., Rempfler, A., Buchholz, J., Hartig, J. & Ulrich-Riedhammer, E. M. (2018). System competence modelling: Theoretical foundation and empirical validation of a model involving natural, social and human-environment systems. *Journal of Research in Science Teaching* 55(5), 685-711. First published online 2017.

Midgley, G. (2000). Systemic Intervention. Springer US (Contemporary Systems Thinking).

Milstein, B., Homer, J., Briss, P., Burton, D. & Pechacek, T. (2011). Why Behavioral And Environmental Interventions Are Needed To Improve Health At Lower Cost. *Health Affairs* 30(5), 823-832.

Mingers, J. (2011). Soft OR comes of age—but not everywhere!. *Omega* 39(6), 729-741.

Mingers, J. & Rosenhead, J. (2004). Problem structuring methods in action. *European Journal of Operational Research* 152(3), 530-554.

Mingers, J. & Taylor, S. (1992). The Use of Soft Systems Methodology in Practice. *Journal of the Operational Research Society* 43(4), 321-332.

Mitchell, M. (2009). Complexity: A Guided Tour. New York: Oxford University Press.

Monat, J.P. & Gannon, T.F. (2015). What is systems thinking? A review of selected literature plus recommendations. American Journal of Systems Science 4(1), 11-26. doi:10.5923/j.ajss.20150401.02.

Moore, G. F., Evans, R. E., Hawkins, J., Littlecott, H., Melendez-Torres, G., Bonell, C. et al. (2019). From complex social interventions to interventions in complex social systems: Future directions and unresolved questions for intervention development and evaluation. *Evaluation* 25(1), 23-45. First published online 2018.

Morecroft, J. D. (1983). System dynamics: Portraying bounded rationality. *Omega* 11(2), 131-142.

Morecroft, J.D.W. & Sterman, J.D. (eds.) (1994). Modeling for Learning Organizations. Portland, OR: Productivity Press.

Morozov, E. (2014, October 13). The planning machine: Project Cybersyn and the origins of the Big Data nation. The New Yorker.

Moxnes, E. (1998). Not Only the Tragedy of the Commons: Misperceptions of Bioeconomics. *Management Science* 44(9), 1234-1248.

Moxnes, E. (2000). Not only the tragedy of the commons: misperceptions of feedback and policies for sustainable development. *System Dynamics Review* 16(4), 325-348.

Moxnes, E. (2004). Misperceptions of basic dynamics: the case of renewable resource management. *System Dynamics Review* 20(2), 139-162.

Moxnes, E. & Saysel, A. K. (2009). Misperceptions of global climate change: information policies. *Climatic Change* 93(1-2), 15-37. First published online 2008.

Munro, E. (2011). The Munro Review of Child Protection: Final Report, A Child-Centred System. Cm 8062. London: Department for Education.

Narayanan, A. & Moritz, B. B. (2015). Decision Making and Cognition in Multi-Echelon Supply Chains: An Experimental Study. *Production and Operations Management* 24(8), 1216-1234.

Nebel, A., Kling, A., Willamowski, R. & Schell, T. (2024). Recalibration of limits to growth: An update of the World3 model. *Journal of Industrial Ecology* 28(1), 87-99. First published online 2023.

Newell, B. R., Kary, A., Moore, C. & Gonzalez, C. (2015). Managing the Budget: Stock-Flow Reasoning and the CO 2 Accumulation Problem. *Topics in Cognitive Science* 8(1), 138-159.

Nordhaus, W. D. (1973). World Dynamics: Measurement Without Data. *The Economic Journal* 83(332), 1156.

Nordhaus, W. D. (1992). Lethal Model 2: The Limits to Growth Revisited. *Brookings Papers on Economic Activity* 1992(2), 1.

Occhipinti, J. A., Skinner, A., Iorfino, F., Lawson, K., Sturgess, J., Burgess, W. et al. (2021). Reducing youth suicide: systems modelling and simulation to guide targeted investments across the determinants. *BMC Medicine* 19(1).

Oliva, R. & Sterman, J. D. (2001). Cutting Corners and Working Overtime: Quality Erosion in the Service Industry. *Management Science* 47(7), 894-914.

Oreskes, N., Shrader-Frechette, K. & Belitz, K. (1994). Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science* 263(5147), 641-646.

Örtenblad, A. (2002). A Typology of the Idea of Learning Organization. *Management Learning* 33(2), 213-230.

Osman, M. (2010). Controlling uncertainty: A review of human behavior in complex dynamic environments. *Psychological Bulletin* 136(1), 65-86.

Ossimitz, G. (2002). Stock-flow-thinking and reading stock-flow-related graphs: An empirical investigation in dynamic thinking abilities. *Proceedings of the 20th International Conference of the System Dynamics Society*, Palermo.

Ostrom, E. (1990). Governing the Commons. Cambridge University Press.

Ostrom, E. (2005). Understanding Institutional Diversity. Princeton: Princeton University Press.

Ostrom, E. (2009). A General Framework for Analyzing Sustainability of Social-Ecological Systems. *Science* 325(5939), 419-422.

Ostrom, E. (2010). Beyond Markets and States: Polycentric Governance of Complex Economic Systems. *American Economic Review* 100(3), 641-672.

Ostrom, E., Janssen, M. A. & Anderies, J. M. (2007). Going beyond panaceas. *Proceedings of the National Academy of Sciences* 104(39), 15176-15178.

Ostrom, E., Walker, J. & Gardner, R. (1992). Covenants with and without a Sword: Self-Governance Is Possible. *American Political Science Review* 86(2), 404-417.

Owen, B., Brown, A. D., Kuhlberg, J., Millar, L., Nichols, M., Economos, C. et al. (2018). Understanding a successful obesity prevention initiative in children under 5 from a systems perspective. *PLOS ONE* 13(3), e0195141.

Paich, M. & Sterman, J. D. (1993). Boom, Bust, and Failures to Learn in Experimental Markets. *Management Science* 39(12), 1439-1458.

Paina, L. & Peters, D. H. (2012). Understanding pathways for scaling up health services through the lens of complex adaptive systems. *Health Policy and Planning* 27(5), 365-373. First published online 2011.

Pala, Ö. & Vennix, J. A. M. (2005). Effect of system dynamics education on systems thinking inventory task performance. *System Dynamics Review* 21(2), 147-172.

Patriarca, R., Di Gravio, G., Woltjer, R., Costantino, F., Praetorius, G., Ferreira, P. et al. (2020). Framing the FRAM: A literature review on the functional resonance analysis method. *Safety Science* 129, 104827.

Pedler, M. & Burgoyne, J. G. (2017). Is the learning organisation still alive?. *The Learning Organization* 24(2), 119-126.

Perrow, C. (1984). Normal Accidents: Living with High-Risk Technologies. New York: Basic Books.

Peters, D. H. (2014). The application of systems thinking in health: why use systems thinking?. *Health Research Policy and Systems* 12(1).

Petticrew, M., Knai, C., Thomas, J., Rehfuess, E. A., Noyes, J., Gerhardus, A. et al. (2019). Implications of a complexity perspective for systematic reviews and guideline development in health decision making. *BMJ Global Health* 4(Suppl 1), e000899.

Pickering, A. (2010). The Cybernetic Brain: Sketches of Another Future. Chicago: University of Chicago Press.

Plate, R. (2010). Assessing individuals' understanding of nonlinear causal structures in complex systems. *System Dynamics Review* 26(1), 19-33.

Plsek, P. E. & Greenhalgh, T. (2001). The challenge of complexity in health care. *BMJ* 323(7313), 625-628.

Public Health England (2019). Whole Systems Approach to Obesity: A Guide to Support Local Approaches to Promoting a Healthy Weight. London: PHE.

Rahmandad, H., Lim, T. Y. & Sterman, J. (2021). Behavioral dynamics of COVID-19: estimating underreporting, multiple waves, and adherence fatigue across 92 nations. *System Dynamics Review* 37(1), 5-31.

Rahmandad, H., Repenning, N. & Sterman, J. (2009). Effects of feedback delay on learning. *System Dynamics Review* 25(4), 309-338.

Rahmandad, H. & Sterman, J. (2008). Heterogeneity and Network Structure in the Dynamics of Diffusion: Comparing Agent-Based and Differential Equation Models. *Management Science* 54(5), 998-1014.

Rahmandad, H., Xu, R. & Ghaffarzadegan, N. (2022). Enhancing long-term forecasting: Learning from COVID-19 models. *PLOS Computational Biology* 18(5), e1010100.

Ramage, M. & Shipp, K. (2020). Systems Thinkers. Springer London.

Randle, J. M. & Stroink, M. L. (2018). The Development and Initial Validation of the Paradigm of Systems Thinking. *Systems Research and Behavioral Science* 35(6), 645-657.

Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. *Safety Science* 27(2-3), 183-213.

Reason, J. (1990). Human Error. Cambridge University Press.

Reason, J. (1997). Managing the Risks of Organizational Accidents. Aldershot: Ashgate.

Reason, J. (2000). Human error: models and management. *BMJ* 320(7237), 768-770.

Repenning, N. P. & Sterman, J. D. (2001). Nobody Ever Gets Credit for Fixing Problems That Never Happened: Creating and Sustaining Process Improvement. *California Management Review* 43(4), 64-88.

Repenning, N. P. & Sterman, J. D. (2002). Capability Traps and Self-Confirming Attribution Errors in the Dynamics of Process Improvement. *Administrative Science Quarterly* 47(2), 265-295.

Reynolds, M. & Holwell, S. (eds.) (2010). Systems Approaches to Managing Change: A Practical Guide. Springer London.

Richardson, G. P. (1986). Problems with causal-loop diagrams. *System Dynamics Review* 2(2), 158-170.

Richardson, G.P. (1991). Feedback Thought in Social Science and Systems Theory. Philadelphia: University of Pennsylvania Press.

Richardson, G. P. (2011). Reflections on the foundations of system dynamics. *System Dynamics Review* 27(3), 219-243.

Richardson, K., Steffen, W., Lucht, W., Bendtsen, J., Cornell, S. E., Donges, J. F. et al. (2023). Earth beyond six of nine planetary boundaries. *Science Advances* 9(37).

Richmond, B. (1993). Systems thinking: Critical thinking skills for the 1990s and beyond. *System Dynamics Review* 9(2), 113-133.

Riess, W. & Mischo, C. (2010). Promoting Systems Thinking through Biology Lessons. *International Journal of Science Education* 32(6), 705-725. First published online 2009.

Rittel, H. W. J. & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences* 4(2), 155-169.

Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, F. S., Lambin, E. F. et al. (2009). A safe operating space for humanity. *Nature* 461(7263), 472-475.

Rosenblueth, A., Wiener, N. & Bigelow, J. (1943). Behavior, Purpose and Teleology. *Philosophy of Science* 10(1), 18-24.

Rosenhead, J. (ed.) (1989). Rational Analysis for a Problematic World. Chichester: Wiley.

Rosenhead, J. & Mingers, J. (eds.) (2001). Rational Analysis for a Problematic World Revisited. Chichester: Wiley.

Rouwette, E. A. J. A., Vennix, J. A. M. & Mullekom, T. v. (2002). Group model building effectiveness: a review of assessment studies. *System Dynamics Review* 18(1), 5-45.

Rusoja, E., Haynie, D., Sievers, J., Mustafee, N., Nelson, F., Reynolds, M. et al. (2018). Thinking about complexity in health: A systematic review of the key systems thinking and complexity ideas in health. *Journal of Evaluation in Clinical Practice* 24(3), 600-606.

Rutter, H., Savona, N., Glonti, K., Bibby, J., Cummins, S., Finegood, D. T. et al. (2017). The need for a complex systems model of evidence for public health. *The Lancet* 390(10112), 2602-2604.

Salmon, P. M., Cornelissen, M. & Trotter, M. J. (2012). Systems-based accident analysis methods: A comparison of Accimap, HFACS, and STAMP. *Safety Science* 50(4), 1158-1170.

Saltelli, A., Bammer, G., Bruno, I., Charters, E., Di Fiore, M., Didier, E. et al. (2020). Five ways to ensure that models serve society: a manifesto. *Nature* 582(7813), 482-484.

Schaffernicht, M. & Groesser, S. N. (2011). A comprehensive method for comparing mental models of dynamic systems. *European Journal of Operational Research* 210(1), 57-67.

Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V. et al. (2009). Early-warning signals for critical transitions. *Nature* 461(7260), 53-59.

Scheffer, M., Carpenter, S., Foley, J. A., Folke, C. & Walker, B. (2001). Catastrophic shifts in ecosystems. *Nature* 413(6856), 591-596.

Scott, J.C. (1998). Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed. New Haven: Yale University Press.

Scott, R. J., Cavana, R. Y. & Cameron, D. (2016). Recent evidence on the effectiveness of group model building. *European Journal of Operational Research* 249(3), 908-918.

Seddon, J. (2008). Systems Thinking in the Public Sector: The Failure of the Reform Regime and a Manifesto for a Better Way. Axminster: Triarchy Press.

Senge, P.M. (1990). The Fifth Discipline: The Art and Practice of the Learning Organization. New York: Doubleday/Currency.

Senge, P.M. et al. (1994). The Fifth Discipline Fieldbook. New York: Currency Doubleday.

Senge, P. M. & Sterman, J. D. (1992). Systems thinking and organizational learning: Acting locally and thinking globally in the organization of the future. *European Journal of Operational Research* 59(1), 137-150.

Simon, H.A. (1962). The architecture of complexity. Proceedings of the American Philosophical Society 106(6), 467-482.

Simon, H.A. (1969). The Sciences of the Artificial. Cambridge, MA: MIT Press.

Skivington, K., Matthews, L., Simpson, S. A., Craig, P., Baird, J., Blazeby, J. M. et al. (2021). A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *BMJ*, n2061.

Snowden, D.J. & Boone, M.E. (2007). A leader's framework for decision making. Harvard Business Review 85(11), 68-76. PMID 18159787.

Solow, R. M. (1973). Is the End of the World at Hand?. *Challenge* 16(1), 39-50.

Spear, S. & Bowen, H. K. (1999). Decoding the DNA of the Toyota Production System. *Harvard Business Review*, September-October 1999.

Stadler, M., Becker, N., Gödker, M., Leutner, D. & Greiff, S. (2015). Complex problem solving and intelligence: A meta-analysis. *Intelligence* 53, 92-101.

Stave, K. & Hopper, M. (2007). What constitutes systems thinking? A proposed taxonomy. Proceedings of the 25th International Conference of the System Dynamics Society, Boston.

Steckel, J. H., Gupta, S. & Banerji, A. (2004). Supply Chain Decision Making: Will Shorter Cycle Times and Shared Point-of-Sale Information Necessarily Help?. *Management Science* 50(4), 458-464.

Sterman, J. (2018). System dynamics at sixty: the path forward. *System Dynamics Review* 34(1-2), 5-47.

Sterman, J. D. (1987). Testing Behavioral Simulation Models by Direct Experiment. *Management Science* 33(12), 1572-1592.

Sterman, J. D. (1989a). Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment. *Management Science* 35(3), 321-339.

Sterman, J. D. (1989b). Misperceptions of feedback in dynamic decision making. *Organizational Behavior and Human Decision Processes* 43(3), 301-335.

Sterman, J.D. (1991). A skeptic's guide to computer models. In G.O. Barney et al. (eds.), Managing a Nation: The Microcomputer Software Catalog (pp. 209-229). Boulder: Westview.

Sterman, J. D. (1994). Learning in and about complex systems. *System Dynamics Review* 10(2-3), 291-330.

Sterman, J.D. (2000). Business Dynamics: Systems Thinking and Modeling for a Complex World. Boston: Irwin/McGraw-Hill.

Sterman, J. D. (2001). System Dynamics Modeling: Tools for Learning in a Complex World. *California Management Review* 43(4), 8-25.

Sterman, J. D. (2002). All models are wrong: reflections on becoming a systems scientist. *System Dynamics Review* 18(4), 501-531.

Sterman, J. D. (2006). Learning from Evidence in a Complex World. *American Journal of Public Health* 96(3), 505-514.

Sterman, J. D. (2008). Risk Communication on Climate: Mental Models and Mass Balance. *Science* 322(5901), 532-533.

Sterman, J. D. (2010). Does formal system dynamics training improve people's understanding of accumulation?. *System Dynamics Review* 26(4), 316-334.

Sterman, J. D. & Dogan, G. (2015). "I'm not hoarding, I'm just stocking up before the hoarders get here.". *Journal of Operations Management* 39-40(1), 6-22.

Sterman, J. D., Repenning, N. P. & Kofman, F. (1997). Unanticipated Side Effects of Successful Quality Programs: Exploring a Paradox of Organizational Improvement. *Management Science* 43(4), 503-521.

Sterman, J. D. & Sweeney, L. B. (2002). Cloudy skies: assessing public understanding of global warming. *System Dynamics Review* 18(2), 207-240.

Sterman, J. D. & Sweeney, L. B. (2007). Understanding public complacency about climate change: adults' mental models of climate change violate conservation of matter. *Climatic Change* 80(3-4), 213-238.

Sterman, J., Fiddaman, T., Franck, T., Jones, A., McCauley, S., Rice, P. et al. (2012). Climate interactive: the C-ROADS climate policy model. *System Dynamics Review* 28(3), 295-305.

Sterman, J., Franck, T., Fiddaman, T., Jones, A., McCauley, S., Rice, P. et al. (2015). World climate. *Simulation & Gaming* 46(3-4), 348-382. First published online 2014.

Stroh, D.P. (2015). Systems Thinking for Social Change. White River Junction, VT: Chelsea Green.

Strohschneider, S. & Güss, D. (1999). The Fate of the Moros: A Cross-cultural Exploration of Strategies in Complex and Dynamic Decision Making. *International Journal of Psychology* 34(4), 235-252.

Sulaman, S. M., Beer, A., Felderer, M. & Höst, M. (2019). Comparison of the FMEA and STPA safety analysis methods–a case study. *Software Quality Journal* 27(1), 349-387. First published online 2017.

Sweeney, L. B. & Sterman, J. D. (2000). Bathtub dynamics: initial results of a systems thinking inventory. *System Dynamics Review* 16(4), 249-286.

Sweeney, L. B. & Sterman, J. D. (2007). Thinking about systems: student and teacher conceptions of natural and social systems. *System Dynamics Review* 23(2-3), 285-311.

Swinburn, B. A., Kraak, V. I., Allender, S., Atkins, V. J., Baker, P. I., Bogard, J. R. et al. (2019). The Global Syndemic of Obesity, Undernutrition, and Climate Change: The Lancet Commission report. *The Lancet* 393(10173), 791-846.

Thibodeau, P. H., Frantz, C. M. & Stroink, M. L. (2016). Situating a Measure of Systems Thinking in a Landscape of Psychological Constructs. *Systems Research and Behavioral Science* 33(6), 753-769. First published online 2015.

Trochim, W. M., Cabrera, D. A., Milstein, B., Gallagher, R. S. & Leischow, S. J. (2006). Practical Challenges of Systems Thinking and Modeling in Public Health. *American Journal of Public Health* 96(3), 538-546.

Turner, G. (2008). A comparison of The Limits to Growth with 30 years of reality. *Global Environmental Change* 18(3), 397-411.

Ulrich, W. (1981). A critique of pure cybernetic reason: The Chilean experience with cybernetics. *Journal of Applied Systems Analysis* 8, 33-59.

Ulrich, W. (1983). Critical Heuristics of Social Planning. Bern: Haupt.

Ulrich, W. (1987). Critical heuristics of social systems design. *European Journal of Operational Research* 31(3), 276-283.

Underwood, P. & Waterson, P. (2013). Systemic accident analysis: Examining the gap between research and practice. *Accident Analysis & Prevention* 55, 154-164.

Underwood, P. & Waterson, P. (2014). Systems thinking, the Swiss Cheese Model and accident analysis: A comparative systemic analysis of the Grayrigg train derailment using the ATSB, AcciMap and STAMP models. *Accident Analysis & Prevention* 68, 75-94.

Vandenbroeck, P., Goossens, J. & Clemens, M. (2007). Tackling Obesities: Future Choices, Building the Obesity System Map. London: Government Office for Science.

Vaughan, D. (1996). The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA. Chicago: University of Chicago Press.

Vennix, J.A.M. (1996). Group Model Building: Facilitating Team Learning Using System Dynamics. Chichester: Wiley.

Villa, S., Gonçalves, P. & Arango, S. (2015). Exploring retailers' ordering decisions under delays. *System Dynamics Review* 31(1-2), 1-27.

von Bertalanffy, L. (1950). An outline of general system theory. *The British Journal for the Philosophy of Science* 1(2), 134-165.

von Bertalanffy, L. (1968). General System Theory: Foundations, Development, Applications. New York: George Braziller.

von Foerster, H. (2003). Understanding Understanding. Springer New York.

Walker, B., Holling, C. S., Carpenter, S. R. & Kinzig, A. P. (2004). Resilience, Adaptability and Transformability in Social-ecological Systems. *Ecology and Society* 9(2).

Waterson, P., Robertson, M. M., Cooke, N. J., Militello, L., Roth, E. & Stanton, N. A. (2015). Defining the methodological challenges and opportunities for an effective science of sociotechnical systems and safety. *Ergonomics* 58(4), 565-599.

Weaver, W. (1948). Science and complexity. *American Scientist* 36(4), 536-544.

Weinhardt, J. M., Hendijani, R., Harman, J. L., Steel, P. & Gonzalez, C. (2015). How analytic reasoning style and global thinking relate to understanding stocks and flows. *Journal of Operations Management* 39-40(1), 23-30.

Wiener, N. (1948). Cybernetics, or Control and Communication in the Animal and the Machine. New York: Wiley / Cambridge, MA: MIT Press.

Wilensky, U. & Resnick, M. (1999). Thinking in Levels: A Dynamic Systems Approach to Making Sense of the World. *Journal of Science Education and Technology* 8(1), 3-19.

Wilson, D. S., Ostrom, E. & Cox, M. E. (2013). Generalizing the core design principles for the efficacy of groups. *Journal of Economic Behavior & Organization* 90, S21-S32.

Wolstenholme, E. F. (2003). Towards the definition and use of a core set of archetypal structures in system dynamics. *System Dynamics Review* 19(1), 7-26.

Woods, D. D. (2015). Four concepts for resilience and the implications for the future of resilience engineering. *Reliability Engineering & System Safety* 141, 5-9.

Wu, D. Y. & Katok, E. (2006). Learning, communication, and the bullwhip effect. *Journal of Operations Management* 24(6), 839-850. First published online 2005.

Wüstenberg, S., Greiff, S. & Funke, J. (2012). Complex problem solving — More than reasoning?. *Intelligence* 40(1), 1-14.

Yang, B., Watkins, K. E. & Marsick, V. J. (2004). The construct of the learning organization: Dimensions, measurement, and validation. *Human Resource Development Quarterly* 15(1), 31-55.

York, S., Lavi, R., Dori, Y. J. & Orgill, M. (2019). Applications of Systems Thinking in STEM Education. *Journal of Chemical Education* 96(12), 2742-2751.