After the Builder **Walks Away**

Machine Intelligence from Stone Age Traps to Conversational AI

Douglas C. Youvan

After the Builder Walks Away

Douglas C. Youvan

Revised expanded edition / November 2026

Developed and written by Douglas C. Youvan in collaboration with ChatGPT. Chapter 12 includes the assistant's account of its role and its recorded replies to four cases. Numbered references identify historical and technical sources. The proposed weather instrument is examined in simulation and against six years of archived observations from Parsons, Kansas.

Contents

Prologue The Second Time Through 1

PART I

What Can an Object Do Alone?

01 The Fisher Goes Home 5

02 The Clock Measures Its Own Water 12

03 Who Moved the Pebble? 21

PART II

When Experience Changes the Machine

04 Two Chess Players in Cabinets 30

05 When the Maze Writes the Memory 38

06 A Machine Searches for Its Balance 47

07 The Position It Has Never Seen 54

Contents continued

PART III

Who Chooses the Next Question?

08 A Conversation and a Room Full of Blocks 61

09 Intelligence at a Distance 68

10 The Scientist That Asks the Yeast 75

11 A Machine Becomes Its Own Opponent 85

12 The Machine Answering You Now 92

PART IV

Back to the Stone Age

13 Three Stones, Three Cords, and a Question 100

14 Two Piles of Stones and the Turing Boundary 106

15 Build It, Start It, Leave It 113

Epilogue What Would an Archaeologist Find? 128

Notes on Sources and Experiments 133

References 138

PROLOGUE

The Second Time Through

There is a moment in Thomas Ross's maze experiment when almost nothing happens. A probe begins to enter a passage. It comes back. On the previous run it went much farther. Someone watching from the side might take the difference for a sticky bearing. In an account of machine intelligence it deserves a closer look. Something happened at the far end of the passage on the first run that can now govern the arm before it gets there.

The apparatus appeared in Scientific American in April 1933. An upright plate had passages cut through it. A movable arm carried a sensitive tip. Switches, electromagnets, sliding contacts, batteries, and clockwork did the work. There were four memory cells. The tip's encounters with the ends of the passages could change which circuit would be completed when the arm was again opposite a particular opening. Returning the arm to its starting point gave the machine another chance at the same surroundings. Its second path differed from its first. Ross included a drawing of a still more direct path that a refined apparatus might take. He distinguished the path his model actually took from the one he thought it could be made to take. The distinction is generous to a reader nearly a century later. We know which machine is on the bench. [1]

The person's hand returning the arm to the beginning matters. It tells us exactly where to place an objection. If the claim is that the machine organized a whole program of experiments, the hand is evidence against it. If the claim is that the machine acquired a different response to a particular passage, the hand explains very little. A person who had never seen the first run could return the arm, and the switch would still be set. Draw a curtain between operator and maze and the result would remain intelligible. Disconnect the contact that writes the memory and a prediction follows. The arm might still move quite impressively, but this particular shortening of its path should disappear.

Consider what an observer is entitled to say about that first run. A passage ends in a sloping surface. Contact bends the probe tongue. A circuit closes.

There is nothing troubling about those verbs. Then a switch retains a changed position. The next movement depends on the position left behind. We might say the machine knows that the passage is blind. Someone else might prefer to say that it has acquired a disposition to withdraw. We can postpone that quarrel. Both descriptions owe the reader the same piece of evidence. The prior encounter must make a difference after the contact that produced it has ceased. Otherwise a word like memory buys us a remarkable performance at no cost to the mechanism.

The title of Ross's article was “Machines That Think.” In 1933, readers did not need to have heard of a digital computer to find that title provocative. The editorial preface treated the mechanical explanation of thought as a hypothesis with philosophical rivals. Ross gave his readers a photograph and a circuit. There is a productive disparity of scale here. An editorial can ask whether the human will is free. A photograph can show whether a particular switch is mounted on a pivot. The first question survives the answer to the second. Yet the two are often run together in stories about intelligent machines. This book will stay long enough with individual switches, water levels, chess positions, faulty instruments, and sentences to find out what work they actually do.

Our investigation began with the imaginary loss of almost everything that makes a modern computer possible. Suppose a community retained its knowledge and found itself with wood, fiber, clay, stone, and water. How directly could it build something that deserved to be called intelligent? A bowl of marked stones seemed a splendid answer. A person could draw one to choose a move and add others of the same kind after a win. Gradually the odds would change. The materials could be assembled before supper. Then came a question with the force of pulling a plug. Who notices the win and adds the stones?

If people are allowed to do those things during operation, there is an enormous and fascinating history of intelligence distributed between people and artifacts. A navigator reading seas that a chart has helped her understand belongs in it. A laboratory technician reading the trace of a delicate instrument belongs in it. A human calculator equipped with an abacus belongs in it. Here we take a different unit of investigation. How much of a particular act of judgment survives when the builder walks away? The person may supply a goal, give an opening position, wind the spring, or offer the question. If a later

step requires a person to determine what the machine should conclude in that case, that step goes on the person's side of the account.

An objection arises at once. Surely the builder put it all there. Ross chose the magnets and wired the contacts. A different slope at the end of a passage might have taught the machine a different response. Yet a person can build a rain gauge and go home ignorant of tomorrow's rainfall. The builder has provided a way for an event to matter. The event itself supplies the information that distinguishes this outcome from another.

Then what of a stone that rolls downhill and turns when it hits a root? Its owner can certainly be elsewhere. The title of this book tells us whose work we may credit. To say what the work amounts to, we need a problem, some alternative answers and a way to tell a good answer from a poor one. Two devices made with equally modest materials will let us ask what else intelligence might require.

Thread three cords through holes in a board and tie their upper ends in a knot. Equal hanging stones pull the knot toward the place that minimizes the combined distance to the holes. Move the holes to represent three new villages and the apparatus can find a site on its first release. Now give a second device a vane, a rain cup and four wooden sliders. Two identical winds weeks apart can produce different forecasts because the rains between them have moved a slider. The cord board solves a fresh problem by keeping its method. The weather instrument can revise its answer by retaining a history.

A reader who requires learning will draw the boundary between those two machines. A reader who asks whether a device can solve an unfamiliar instance unaided may admit both. Those choices have different experimental consequences. Shift a hole and measure the knot. Erase a slider and watch the flag. Either reader can tell us in advance what would count against the claim. That is the sense in which a history of machine intelligence can begin with a fisher walking away.

The machines in these pages were never entered in one contest. A lamp in medieval Baghdad and a spacecraft carrying artificial intelligence into space

faced very different economies of failure. A lamp that quietly tended its wick conserved an attendant's attention. A spacecraft unable to revise a plan could postpone work until a ground crew had diagnosed a problem far from the scene. A spoken answer can save the time of a reader or consume much more of it if its factual basis is obscure. A history ordered only by increasing cleverness misses the practical reasons for handing a particular function to a device. Often the first reason was quite modest. The fisher had somewhere else to go.

Near the end we shall return to the imagined community and give it a less obliging instructor than a person with a supply of pebbles. We shall let the wind pose a question and rain settle the outcome. The mechanical details will matter as much as the idea. If a wind vane swings during a storm and changes which memory is updated, the machine may appear to predict rain by recording what the air is doing after rain has already arrived. A single latch separates a defensible forecast from a retrospective account. The design remains a proposal. Its usefulness can be addressed with measurements and its historical status can be stated in one sentence: no ancient instance of this device is being claimed.

For the moment keep one scene in mind. A person has returned an arm to its starting point. The current is on. At the entrance to a passage the arm begins to withdraw. By looking carefully at the separate contributions of the operator, the environment, and the memory cell, we can account for something a bystander would reasonably call a lesson learned. The smallness of the lesson is one of its virtues. We know where to look for it.

CHAPTER 1

The Fisher Goes Home

The five hazel traps lay beneath a modern city, some 6.3 meters below mean sea level in the Liffey estuary. The excavators, Melanie McQuade and Lorna O'Donnell, dated the activity to about 6100–5700 calibrated BCE. Woven wood survived well enough for them to examine how it had been worked. From the material they also inferred the management of hazel on a coppicing cycle of about eight years. A reader looking for cleverness in the recovered objects has several places to look before reaching for the word intelligence. There is knowledge of the growth of a tree. There is a skill of the fingers. There is an understanding of a river that changes direction with the tide. [2]

What the excavators recovered was more fortunate than a drawing in one respect and poorer in another. They could examine real rods handled by real people. They could not ask the people how a promising place was recognized, how often the catch was inspected, or what might induce a fisher to shift the trap a little farther along the bank. Waterlogged wood has preserved part of the operation by a process very different from a writer's choice of what to record. The distinction between what survives and what was done is going to matter whenever we seek ancestors for a machine.

The work contained in a shape

Imagine an ordinary funnel trap. This is an explanation of the principle, independent of the particular form of each Liffey find. A fish swimming along a barrier encounters an opening. The entrance guides it inward. To leave it has to find a smaller opening from a less helpful direction. The fish can be a perfectly competent swimmer and still spend a long time inside. The mechanism exploits an asymmetry in encounters. A wide collecting area leads to a narrow point in one direction. On the return journey many paths lead to the wall. The fisher's interesting achievement is to have made chance encounters unequal. The fish supplies much of the motion.

A person could achieve a similar practical result by watching an entrance and closing a gate at the right instant. The trap dispenses with the watcher's

recognition and timing. It has no occasion to identify a species, estimate the market value of a catch, or anticipate dinner. A branch of the right size might enter too. If small and large fish pass through differently, that selectivity can be described and tested in physical terms. The shape has inherited some of the work of selection. Anyone who has watched a badly placed funnel collect drifting vegetation will appreciate just how incomplete the inheritance can be.

The interesting comparison is with a basket standing on dry ground. Quite similar weave, quite different work. An archaeologist identifying a fish trap has to bring together the orientation of its parts, evidence about water, context of deposition, and parallels with other objects. An engineer identifying a device that learns must do something analogous. A rack of sliding weights looks like a splendid store of memories until it turns out that a technician sets the weights once a year to keep a structure level. A slot suggests that something can be moved. The slot alone cannot establish who did the moving, how often it happened, or which observation prompted it. We shall learn to demand quite a lot from evidence without disparaging what a well-preserved object really can tell us.

For the fisher the test is more immediate. Go away for an interval and return. Was anything caught? A successful catch demonstrates that presence at the moment of capture was unnecessary. The same observation says little about whether the trap can improve itself. Ten fish in the second interval after two in the first could be the result of a changing tide, a passing school, or an improvement in the trap. A full basket measures an outcome produced jointly by the world and the device. It does not allocate credit between them. In modern evaluations a higher score after another day of deployment sometimes receives a similarly hasty interpretation. The difficulty is old enough to be met at the mouth of a river.

A person can be absent in several ways

Imagine three versions of our fisher's work. In one the person sets the trap, waits on the bank, and retrieves a catch, leaving the arrangement as it was. In another the person departs before the incoming tide. In a third the fisher returns after every tide, observes where fish have gathered, and adjusts a movable wing of the trap. In all three a future archaeologist might recover virtually the same wood. Yet the third operation has an adaptive element that the first two, as described, lack. The adjustment depends on an outcome and can change the chance of future capture. Its author is a person. The absence of

the fisher from the moment of capture never transferred responsibility for that adjustment.

Now introduce a fourth version. An incoming surge pushes a wing into a new position. The next catch changes because the water itself changed the device. This is a memory in the weak physical sense in which a bent twig remembers a load. Whether the change constitutes improvement remains to be shown. It could send every subsequent fish into open water. The wind can bend an antenna in a direction that improves reception for an afternoon. To promote the event to learning, one would want evidence that the changes occur in a way that tends to exploit past experience under a defined set of conditions. Otherwise a lucky accident has acquired the reputation of a competent operator.

A fifth version brings out a possibility the original trap lacks. Suppose each of two entrances has a small flap and the passage of fish through one shifts a balance that gradually favors that entrance. The organization of tomorrow's trap would then depend on traffic through today's. It could also be a poor trap. A large fish could jam the favored entrance, or reinforcing early traffic could make the device neglect a newly productive route. Nothing in the definition guarantees successful performance. What changes in this imaginary construction is testability. We could ask whether experience with traffic on the left, followed by identical conditions, produces a different choice from experience with traffic on the right. We could pin the balance and see whether any advantage disappears. We have specified what would have to be added before we gave an adaptive mechanism any share of the credit.

There is no suggestion that this fifth version existed on the Liffey. Developing it in the imagination brings out a logical difference that holds for actual archaeological remains. The further we travel from the recovered wood toward a detailed account of changing its operation, the more independent evidence we need. A possibility that makes good engineering is allowed to remain an invention in a book of history. Confusing a reconstruction that works with a demonstration that ancient people built it would close off much of what is interesting about both design and historical inquiry.

The tide supplies power and information

The tide has two roles in an unattended fishing system. It moves water and living things past a barrier. It also makes certain arrangements more promising at some times than at others. To use the first role, the builder need

only place something where water will move it. To use the second, the arrangement has to distinguish among conditions that matter. An observer who sees water turn a paddle has identified a source of energy. An observer who sees water at a certain height release a shutter has found something more. The height participates in answering a question about when to act. That distinction will return in a weather machine where one source of water drives the clock and a different cup of water determines what is to be remembered.

It is tempting to locate all the intelligence somewhere in the wood because the object has a clear boundary. Even in a simple trap the boundary is a choice. One can remove the basket from the estuary and display it under light. It remains the same archaeological object. Its capacity to catch fish has changed dramatically. The operative system included the route by which the tide brought fish, the tendencies of fish in moving water, and a period in which captives remained available for retrieval. None of those facts makes the fish or the sea a hidden human operator. It means that putting a task into a device often consists in discovering an arrangement in which the rest of the world will do a predictable share of the work.

Ecology also sets the time scale on which an apparently modest device pays off. A tool used at the instant of capture saves effort at that instant. A trap useful for a sequence of hours frees a different resource. Many hours of watching can be exchanged for a period of preparation and brief inspections. If one member of a community knows how to weave the apparatus, another can retrieve its catch. That is a change in how work can be distributed across time and among people. One need not invent a story about the division of labor in Mesolithic Dublin to see the possibility. The absence of the expert during repeated operation is one of the economic motives that will recur in textile machinery, industrial regulation, laboratory automation, and conversational software.

Sharing also separates the preservation of an object from the preservation of a skill. Keep the trap and lose every competent maker, and for a while there may still be fish. When the wood breaks, the situation changes. Keep the community of makers and lose the trap in a flood, and the loss can be repaired. Keep only a drawing of the weave, without knowing where in a tidal channel it belongs, and a fine reconstruction can fail. These cases warn against a familiar shortcut in much later discussions. If a device stops working when the infrastructure that sustains it disappears, it does not follow that a human supplied every answer it gave while that infrastructure was present.

Dependence for manufacture, repair, energy, calibration, and judgment are distinguishable kinds of dependence. The important one depends on the claim being made.

Two ways to leave knowledge behind

An instructive companion to a fish trap is a Marshallese stick chart. The comparison concerns two accomplishments of very different communities. It makes no claim of a common date or historical descent. In the Marshall Islands, navigators used arrangements of sticks and shells to study islands and the behavior of ocean swells. Examples could be studied and memorized ashore. The navigator's trained perception was needed at sea, where the chart itself might be absent from the canoe. The collection of an artifact by a museum can thus preserve the member of a navigational partnership that stayed at home. A viewer who assumes the chart was routinely held in the navigator's hand at every turning point has already assigned the wrong role to it. [27]

Imagine a day with two departures. A fisher leaves a trap to operate in an estuary. A navigator leaves a teaching chart ashore and sets out. In one case the person departs and an environmental interaction continues at the artifact. In the other case a person takes away a capacity developed through earlier interaction with an artifact. The skilled act at sea now runs through body, canoe, and swell. Both departures show that knowledge can travel in time and survive changes of company. They are almost mirror images in the location of the judgment made after departure. Calling the chart a navigation computer would say very little until we decided whether the word referred to the physical object or to an organized practice.

It is easy to undervalue such a chart if one imagines that replacing a person is the natural destiny of every clever device. A teaching object need not replace anyone. It can prepare an observer to notice a distinction that previously passed unnoticed. In an environment where there is no convenient table or surface for consulting diagrams, making a person's own perception more discriminating may be a much better design choice than constructing a bulky aid to consult en route. Measuring the chart by how much navigation it performs unattended would miss the point of its design. Our operational criterion is a way of writing a certain history, never a measure of the value or sophistication of all technologies.

There is a corresponding mistake in assessing a modern training simulator. Suppose pilots become much better at recognizing a particular instrument failure after repeated sessions. The simulator has done a remarkable job even if it is powerless to diagnose anything on an actual aircraft. A separate automatic diagnostic system could be exceptionally good at the same discrimination on board yet teach very little to the crew. Given limited resources, a builder might prefer one, the other, or a combination. The two designs relocate expertise in opposite directions. Historical comparisons give us a language for praising both without silently treating every important capacity as a property of a piece of equipment.

There is a practical consequence for archaeological investigation. If a practice regularly sends its most important symbolic device ashore, the place where one finds the object may be far from the setting in which its lessons were applied. Conversely the findspot of a fish trap can be extraordinarily informative about its operation if the remains are in an appropriate context. Rebuilding an imagined machinery around an object merely because multiple symbols are present can be a mistake about geography as well as about computation. A collection can contain a textbook for a lost sense, a component of an operating device, or the debris of a workshop in which devices were constructed. Those alternatives make quite different predictions about associated traces.

Suppose a better document surfaces from a community for which such judgments are living knowledge. It may tell a museum which curved elements express a relationship of swell and land. That discovery can elevate an object from an apparently decorative arrangement to evidence of sophisticated modeling without converting it into an autonomous navigator. Recognizing intelligence previously overlooked may therefore require giving more credit to a person whose role an exhibit has hidden. In the cases to come it will sometimes require giving more credit to a device. The distribution of credit can move in either direction when evidence improves.

What a trace can establish

Suppose a collection includes a woven object and a row of pegs near it. We have been told that this was a mechanism for changing course according to the last few catches. Before we publish the claim, what could help? Evidence that the pegs occupied more than one position in service would help establish mobility. Wear showing that a peg moved repeatedly in one direction might

identify how a load was applied. Nearby features that linked the catch to the pegs would help identify a possible source of actuation. A duplicate with a complementary arrangement could make a chance coincidence less likely. A reliable description of comparable use might supply the missing link between motion and purpose. Each kind of evidence answers a different part of the proposed history. Some combination might be persuasive. A clever modern reconstruction alone would establish feasibility.

There are pitfalls even in this careful way of looking. An artifact may have had more than one use. Parts of different date may lie next to one another. A modification that makes no sense to someone interested in catching fish might be entirely sensible to someone repairing it with the only materials at hand. A wooden peg polished by many hands can look much like a peg moved by a cord. Knowledge that an object stood in water suggests certain possibilities and rules out others, but the water that preserves wood can also rearrange it. The work of choosing between explanations is an exercise in inference. That work belongs to the archaeological investigators, never retrospectively to the trap.

The same caution can make an exhibit in a modern museum more intelligible. Consider an intact machine whose motor can still be turned on. Its motions may be easier to recover than the intended division of labor. An operator's card could have specified that one particular warning light meant to disregard a trial and reset the system. A film could show the striking movement and cut away before the reset. A photograph could omit an instrument table just outside the frame. In assessing early learning machines we will prize the sources that show the path of an ordinary run, a changed run, and an unsuccessful run. That preference begins with an appreciation of how much future observers would misread if they had only the trap frame.

A wet hazel rod carrying the marks of skilled hands is ample evidence of human ingenuity. There is no reason to turn it into a primitive brain in order to accord it respect. Its contribution to this history is more exact. Here is an object whose function becomes particularly interesting when a person leaves. The object can operate on a new occasion that none of its makers specifically foresaw. Yet if a bend in the river ceases to bring fish, the surviving arrangement supplies no evidence that it can recognize its own failure. For that we must look for a device in which the state of the thing being controlled can travel back, physically, to change the controlling action. Such a path can be made with a float.

CHAPTER 2

The Clock Measures Its Own Water

A hole at the bottom of a full vessel gives time away too quickly. Water above the hole exerts pressure, and as the level falls the pressure falls with it. A scale along the vessel can be marked to allow for the variation, if the outlet and vessel remain sufficiently alike from day to day. Another approach is to give the timing outlet a separate supply whose surface stays at nearly the same height. A solution to a difficulty in the measurement of time becomes a question about controlling a piece of the world. How can an apparatus arrange, as its supply changes, for its immediate surroundings to remain much the same?

Otto Mayr's history of feedback mechanisms, drawing on the account in Vitruvius, describes a float regulator in a water clock of Ctesibius of Alexandria in the first half of the third century BCE. The evidence is a historical report with a chain of transmission. We do not have an operating original with which to settle every detail of construction. We do have a clear physical principle. A float rises with the surface. Its rise restricts the supply. A fall admits more. The level seen by the timing outlet can remain approximately constant. The water whose height the mechanism controls also moves the member that regulates its arrival. [3]

An error can move a valve

A useful way to examine a feedback mechanism is to come at it sideways. Someone quietly increases the outflow from the regulated vessel. At the old inlet setting the surface begins to fall. A falling float permits more supply to enter. If the new inflow compensates for the increased loss before the float reaches the end of its travel, the level settles. Depending on the geometry it may settle somewhat below the old height. Correcting a disturbance need not mean canceling every trace of it. The observer who measures only whether water still comes out misses most of what the apparatus has accomplished.

Now tie the float to its usual position and adjust the inlet by hand until the level is exactly right. Before a disturbance the tied arrangement might give an even prettier demonstration than the free one. Its surface can be perfectly smooth. Increase the outflow again and walk away. The water has lost the route by which its fall could induce an increase in its supply. The old exact adjustment no longer suffices. This is an important experimental device that requires no sophisticated statistics. Two systems can behave alike on a familiar occasion and reveal different organizations when something changes. An excellent performance at the operating point tells us very little about a machine's ability to stay there.

Think of the water flowing through the inlet and the information passing through the float as two separate transactions. In an ordinary installation the float needs only enough force to move a valve or a linkage. Much more water can pass through the opened valve than the float itself displaces. It is the supply that provides the energy. A low-energy change can determine which high-energy process occurs. This practical separation goes a long way toward explaining the appeal of automatic control. It allows an engineer to arrange for a small, easily measured discrepancy to decide the fate of a substantial flow.

There is no need for the mechanism to carry a numerical inscription for its intended level. The effective target can be encoded in a hinge, the length of a lever, and the point where an opening starts to close. Move the hinge and the apparatus tends to another level. The designer retains a say in the goal even while the water decides how much action the discrepancy calls for on each occasion. Language such as “the float compares the level with its target” can be translated back into these physical relations. That translation is especially important when the language of purpose is the first thing a reader sees. There are many ways to represent a goal besides writing it as a number in the memory of a processor.

An elegant mechanism can also have quite prosaic limits. A vessel receiving too little water cannot maintain a level above its available supply. A jammed float can leave the inlet open and turn an apparently intelligent clock into a rather predictable overflow. A clogged time outlet can give an extremely stable level and a poor measure of elapsed time. The machine regulates the variable to which its sensor is coupled. Its builder cares about the inference that variable supports. This gap between a measured quantity and the reason for measuring it will follow us all the way to the automated laboratory. A machine

can regulate one thing admirably while the larger purpose quietly slips out of reach.

Who tends the lamp?

In ninth-century Baghdad the Banū Mūsā, three brothers, compiled the Book of Ingenious Devices. Among its designs is an oil lamp with an arrangement for replenishing oil and advancing a wick. In the mechanism described in Donald Hill's translation and illustrated by the Smithsonian's Lemelson Center, a float falls as oil in a reservoir is consumed. Its movement operates a chain, weight, pulley, wheel, and toothed rod, which advances the wick. A passage from the description can be understood by tracing the members one by one with a finger. The object belongs to the documentary record as a design. Its pages do not tell us how many evenings an individual lamp ran without attention. [4]

This time consider the attendant's former job. A person can look at a flame, notice that it is growing feeble, and draw up a little more wick. In the automatic design the lamp uses a proxy. Consumption of oil induces a movement related to the expected consumption of wick. The little train of mechanisms makes it unnecessary for a person to observe each change. But it does not follow that the lamp observes brightness. If a dark shade screens an otherwise steady flame, the recorded design gives us no reason to expect a compensating movement of the wick. The difference between measuring the outcome and measuring a quantity associated with the outcome matters. A modern engineer might deliberately choose the proxy because it is inexpensive, sturdy, or easy to measure. The choice still places a limit on the range of disturbances the mechanism can correct.

One could put two copies of the imagined lamp in cases. Beside one put a diagram labeled brightness control. Beside the other describe the declining oil level, pulley, and rod. The first label gives a visitor an intention. The second gives the visitor a means of telling which surprises would matter. Oil consumption and flame height may ordinarily correlate. A change of oil or wick can change the relationship. A machine that directly detects brightness might then act quite differently from one indexing wick advance to fuel consumption. The historical interest of the lamp lies partly in the ingenuity with which a quantity available inside the device was made to stand for a more troublesome one. Praise is more informative when it attaches to the mechanism that was actually described.

Once the lamp can tend an aspect of its own operation, the attendant's work changes shape. Filling a reservoir before an evening's use is different from judging every few minutes whether the wick should move. A later museum might describe the device as self-tending and a skeptical visitor might object that it still needs oil. Both can be right about the facts and talking past one another. Self-tending has a scope. Fuel can be supplied without the supplier making the particular adjustment under discussion. In an automated biological laboratory someone may replenish supplies without choosing which hypothesis the next trial tests. The pertinent question concerns the type of contribution, along with its frequency.

A second windmill at right angles

The history of windmills offers a useful mechanical detour on the way from clocks to governors. A mill whose main sails need to face changing winds imposes a job on its keeper. Edmund Lee's eighteenth-century fantail arrangement used a smaller set of vanes oriented at an angle to the main sails to turn the mill's cap into the wind. Once appropriately aligned, the smaller wheel has little reason to keep turning the cap. The very wind that makes alignment valuable can supply both evidence of misalignment and the power to correct it. Compare that economy with an instrument that measures the direction, reports it to the miller, and waits for the miller to act. [28]

A little exercise in geometry brings out what this device can avoid calculating. Attach a vane and a compass rose to a roof. The observer can read a direction relative to north, compare it with the orientation of the main sails, determine which way to turn, and estimate how far. A properly organized crosswind sensor can instead directly convert misalignment into a corrective rotation. It need not represent absolute north at all. Move the compass indications around and the mill can still face the wind, because the relation that matters is between the current air motion and the orientation of the sail assembly. An instrument whose service depends on national map coordinates would have a different requirement. Leaving a quantity out of the representation is often a sign of sound engineering.

The arrangement is also an answer to an objection about primitive components. A small rotation under a modest wind can drive a larger load through gearing, slowly accumulating the required turn. Speed of response and available force trade off. A rapid shift in wind can produce a period of

poor alignment. Unreliable bearings can prevent the small wheel from correcting a small deviation while allowing it to act during a stronger wind. One can describe these limits using the relation between wind, orientation, and correction without assuming that any mill builder carried out a modern mathematical analysis. An invention can embody a successful relationship long before engineers write its equation.

What experience has last week's wind left in an ordinary fantail? The cap's present orientation can preserve the consequence of the last substantial wind. This is a kind of state. If today's wind comes from a new direction the cap has to start somewhere, and yesterday partly decided where. It does not follow that repeated shifts from the west have taught the mill to anticipate the next one. To support that stronger claim, a remembered history would have to alter the response beyond simply leaving the cap in the position it already occupies. One could bring two mills to the same initial orientation after different historical sequences and compare their responses. The idea of matching current physical conditions while changing the past is a powerful test that can be carried from a mechanical roof to a neural network.

In our later rain predictor, an absolute or at least consistently indexed direction does matter. A machine can use the instantaneous relation between wind and its own vane to orient the vane. To remember how often an easterly wind preceded rain, it must repeatedly assign easterly observations to the same memory. If the whole instrument slowly rotates relative to its site, the familiar labels lose their correspondence. The two devices solve nearby problems with very different needs for a stable reference. This is one reason the route from a venerable mechanism to an intelligent weather instrument requires more than combining a mill, a cup, and some counters in a handsome drawing.

For an observer interested in the long history of artificial intelligence, the windmill makes one further point. A device can perform admirably while explicitly representing almost nothing about the outside world. Favorable geometry can make the world perform a transformation the builder would otherwise have had to encode. To say that feedback has saved computation can be a precise engineering statement about this difference. To say that the mill thereby transcends the limits of computation would be a leap into another subject. The same economy of direct coupling makes the paper design for the

rain instrument plausible enough to examine, though remembering an association across a future interval will require a different kind of link.

The engine on a working day

A rotative steam engine built by Boulton and Watt in 1788 survives in the Science Museum Group collection. The museum identifies it as the first such engine to be fitted with a centrifugal governor. The principle can be followed without a boiler. Spinning weights move outward as rotational speed increases. A linkage can use that motion to reduce admission of steam. If the engine slows, the weights fall inward and admission can increase. The final action affects the speed that will subsequently place the weights. [5]

Picture a machine that an engine drives taking on a heavier load. For a moment more torque is required than the engine had been delivering. Its speed begins to drop. If a person were at the throttle, that drop could be a cue to open the valve. The governor conveys a measure of the drop into the motion of the throttle, releasing the person from repeated judgments about how much steam to admit. This comparison gives us a narrow definition of its competence. It does not decide whether the workshop ought to take on the load at all. A machine that chooses a good speed for a new job has inherited a different decision from a machine that holds a speed chosen in advance.

The word governor makes the temptation to overclaim quite strong. We can attribute almost any kind of administrative intelligence to a name. Watch the weights during a disturbance, however, and a more particular difficulty appears. Inertia gives the machine a history. Even after the valve changes position, the engine takes time to respond. If the correction is very strong or delayed, speed can travel too far and a correction in the other direction can follow. A controller that is slow to notice change allows the load to have its way for longer. A controller that responds sharply to every small change can produce oscillation. This is why the study of regulation is richer than the observation that one member moves when another moves. The timing and magnitude of the response help decide whether the loop actually governs anything well.

A numerical illustration makes the issue plain. Imagine a tank in which one unit of a valve setting changes inflow by one liter per minute. A leak suddenly increases by two liters per minute. A valve that opens by one unit for each centimeter of level loss may need about two centimeters of sustained error to

make up the two liters. The precise behavior will depend on pressure and geometry. The figures here define an imagined linear example. The error is the force behind the correction. If a second mechanism slowly changes the resting valve setting while an error persists, a smaller final discrepancy may suffice. That added mechanism retains a little of the history of error. An engineer would recognize integral control in the idea. Before we call it learning, we would want to say what it remembers, what it can change, and whether that change is useful when conditions recur.

This distinction can be tested by removing the extra leak, letting the apparatus settle, and then adding it again. Does the second response anticipate the leak or does it simply start the same recovery anew? A transient can carry information from a few moments ago without storing anything useful for a future encounter. Conversely a controller can store a calibration learned during one period and use it long afterward. There is no timeless frontier at which regulation magically becomes learning. There are specific differences in what prior events can still affect, how long the effects last, and how widely they apply. Those differences can be measured on actual equipment.

The earthquake nobody had to witness

A ball is lying in the mouth of a bronze toad. There are other toads around the vessel and their mouths are empty. Someone walking into the room after an earthquake could in principle read the instrument from that one difference. In the account of Zhang Heng's seismoscope in Han China, conventionally dated to 132 CE, eight dragons around a bronze vessel held balls above corresponding receivers. Earth motion could release a ball. The selected receiver was intended to indicate a direction. The object itself is lost. The historical description preserves much more of its outward appearance than of the mechanism inside. A cutaway on display today shows a proposed way to make it work. [30, 31]

A float in the water clock follows a surface that is present to be followed. The ball can report on motion that has already ceased. Imagine returning the room and the ground as nearly as possible to their former stillness. Something about the event is still available to a reader. A fall under gravity carried a fleeting displacement into a state that could endure for hours. There was no need for an observer to stand beside the instrument and choose a receiver

when the ground moved. The observer arriving later supplies the words “this direction.” The physical choice among outcomes has already happened.

What did the instrument remember? If one of eight initially loaded positions was released, the selection identified a sector. It could not by itself give the hour at which the ground moved, the distance to the source, or a history of the waves. Suppose the same sector were triggered twice before anyone reloaded it. An empty dragon's mouth has no second ball to release. That elementary limitation shows the price paid for keeping a readable trace. An attendant who reloads the balls between events provides maintenance and rearms the instrument. An attendant who feels the shock and drops a ball into the receiver of their own choosing would provide the very discrimination we hoped to attribute to the mechanism.

Memory here means that a past event can still be read. Learning would require a further connection. Imagine a device that compared successive indications with a separate dependable direction measurement, then altered its calibration in response to the differences. Nothing in the cited evidence credits Zhang's instrument with that operation. Even a perfect register of the last hundred earthquakes would leave the question of learning open. What matters is whether one recorded event changes how some future event will be handled.

The lost interior makes this instrument especially instructive. Mechanical engineers Hong-Sen Yan and Kuo-Hung Hsiao found multiple linkages consistent with their interpretation of the historical constraints. A working reconstruction can establish that a proposed linkage could release the ball. Working alone cannot establish which linkage stood inside the ancient vessel, or how reliably the ancient device indicated the source direction under different patterns of ground motion. A single surviving ball and receiver would leave even more room for interpretation. The builder can leave before the instrument registers a shock. Centuries later, when the builder and the vessel are both gone, even a working copy can leave unanswered how the original made its choice. [31]

An intelligence with a specified verb

It is worth standing among these three devices for a moment before racing ahead to a computer. A water clock can regulate a level in a place where a human observer could in principle notice the same changes. A lamp can use

one material change to compensate for another. An engine governor can correct a speed change before a distant owner ever learns that a new load has been connected. All have inherited work from somebody. Only some of that work would ordinarily be described as judgment. If intelligence is to be more than a courtesy title, we should name the action with the precision we expect from a mechanic called to fix it.

Suppose a vendor advertises an intelligent pump. The buyer has questions. Does it stop when a tank is full? Can it distinguish a full tank from a blocked pipe? Would it change its estimate of the time required to fill a tank after repeated trials? Can it choose which of several tanks should be filled before a shortage? Could it discover that a level indicator is faulty and switch to another way of measuring fullness? Each claim suggests a different experiment. Each may require a different kind of evidence and memory. A single number for how smart the pump is would disguise the information most relevant to buying it. Historical comparisons benefit from the same discipline.

There is also something to learn from a deliberate wrong connection. A float that opens an inlet wider as the surface rises will magnify a disturbance. More water raises the float, which admits still more water. A rule for reinforcing a promising move in a game is organized in this second direction. Experience with a particular outcome makes a particular action more likely. Such positive feedback can concentrate effort on a successful move, or it can amplify an early accident. A governor uses opposition to a deviation. A rudimentary learner can use reinforcement of a choice. Both can be called feedback, although interchanging them might respectively flood the clock or prevent a good move from becoming popular. The sign of the connection has consequences.

On returning to the clock, we can ask it a question that the governor and lamp also face. If the clock has run beautifully for a year, which fact has the year taught it? The answer might be none. Wear might even have made it worse. The maker or keeper, meanwhile, may have learned to polish the inlet, change the float, or move it away from a draft. The improvement observed by customers after a year could be entirely real while residing in the keeper's decisions. Our next apparatus looks much more like a learner. To find its corresponding keeper we need only watch what happens at the end of a game.

CHAPTER 3

Who Moved the Pebble?

Set seven stones in a row. Two players will take turns removing one or two. Taking the last stone wins. Now make seven hollows, one for each possible number of stones remaining. Put markers for taking one and taking two in each applicable hollow. To obtain a move from the machine, draw a marker without looking from the hollow for the current position. When only one stone remains there should be only a marker for taking one. Set aside each drawn marker with its hollow so that the moves can be traced afterward. All the materials can be found in a small place. The most conspicuous component still missing from the inventory is a pair of hands.

Suppose the odds begin equal at two stones. Half the time the machine takes both and wins. Half the time it leaves one and sees the other player win. After a win, restore the chosen markers to their respective places and add one of each chosen kind. After a loss, restore them but take away one of each chosen kind, retaining at least one legal marker of each kind. There are many possible reinforcement rules. This one is specified so its consequences can be examined. If a winning move at two is rewarded repeatedly, the hollow for two will come to contain more take-two markers. Anyone can draw from it. The person drawing can be wholly ignorant of how to win.

Donald Michie's MENACE, described in his 1963 paper, used matchboxes for positions in noughts and crosses and colored beads for possible moves. A record of the choices made in a game enabled changes to the bead populations in light of its outcome. Michie subsequently examined related behavior in computer simulations. Our seven-stone game is an illustration devised for this book. Its smaller number of states makes it easier to locate its unacknowledged operator. The relevant historical fact is how little physical equipment it took to embody a changing distribution over choices. [6]

Knowing where the answer is

The little game permits exact analysis. With one stone, take it. With two, take both. With three, either removal gives the other player one of those two winning positions. With four, leave three. With five, leave three again. With six, one has to give the opponent four or five, and a good opponent will know what to do. With seven, take one. This backward reasoning requires no draws and no changes of weights. A person could do it before making a single hollow. A strategy is now available for judging what a learner has acquired. It also gives a teacher an opportunity to influence the results while appearing merely to operate the machine.

At five, taking two is strategically sound against a perfect opponent. If the opponent often makes mistakes, taking one can sometimes work too. Suppose the machine takes one and the opponent, facing four, takes two. The machine wins with two left. Our rule gives the statistically lucky earlier move some credit. The practical question is what one wants to learn: play that exploits a particular opponent, good play against a range of opponents, or a strategy that guarantees a win whenever one is available. Those goals coincide at the two-stone position and diverge at others. An outcome can be objectively observable while the right way to assign credit for the outcome remains debatable.

Drawing at random is essential to this design. A newcomer's willingness to try both moves protects the hollow against a person who always picks the same marker by feel. If the take-two markers happen to be larger, blindfolding the operator may be insufficient. If a heavily reinforced habit crowds out an alternative, exploration becomes rare. Retaining at least one marker helps keep alternatives possible, yet with a hundred winning markers one remaining exploratory marker has a small chance of being drawn. After a change in the opponent's behavior, the very record that made the machine effective can slow its adjustment. We can discuss memory and forgetting without claiming that a hollow has any feelings about a defeat.

Choose a single position and count what one reinforcement does. At five a hollow with three take-one and three take-two markers assigns probability one half to each. An additional take-two marker changes the second probability to four sevenths. A hollow with thirty of each changes to thirty-one sixty-firsts under the same rule. The populations express nearly the same

preference beforehand and react very differently afterward. The ratio tells us about the next draw. The total tells us something about how resistant that preference is to change. Two machines that give identical answers today can have markedly different histories and different futures. A good evaluation sometimes needs to know more than their present output.

The evening shift

Now walk through the work of an actual round. Someone arranges the stones at the start, knows whose turn it is, recognizes how many remain, selects the indicated hollow, makes a fair draw, removes the corresponding number, and records the choice. At the end somebody recognizes which side won. Only then can the listed changes be made at the correct set of hollows. The random draw implements selection according to stored preferences. The human may implement almost everything else. It would be unfair to dismiss the whole thing as a trick. The updated population is a real physical fact with real effects on subsequent play. It would be equally unfair to credit the whole learning loop to the stones.

Replace the operator with a person who cannot play well. The preferences survive. This is a good test of where the preferences reside, and a poor test of whether the device trains itself. Replace the operator with a person who moves every marker accurately but refuses to tell anyone who won. The game can be played indefinitely while the populations remain fixed. Replace the scorer with someone who reports a win whenever the human player wins. The populations change, and their changes systematically support the wrong objective. It is possible for a component that weighs almost nothing, makes hardly any moves, and contributes one bit per game to determine the meaning of all the stored experience.

A counterfactual experiment makes the point sharper. Give two identical pebble machines exactly the same sequence of moves. In the first, reward the winner. In the second, reverse the terminal labels. The physical paths of the stones across the board can be identical up to the instant the updates are made. Watch the later distributions diverge. The lesson was never implicit in the sequence of positions alone. Reaching a board with no stones left means success or failure relative to who took the last stone and which player the apparatus is meant to favor. Credit assignment requires enough information

about the relation between events. A sign reading “last stone wins” is of no use to a latch that cannot distinguish whose turn it was.

This identifies an improvement to the construction. Put a two-position toggle at the turn changer and a follower at the end of a track carrying the playing stones. A move that brings the follower to zero while the toggle points to the machine could release a winning-weight mechanism. The same move at the opposite toggle position could release a different catch. The replacement sounds simple until the mechanical order of events is specified. If the toggle changes before the follower is read, every genuine win will be reported to the wrong side. A device has to preserve the old identity of the mover until the terminal event has been assigned. Timing is part of the logic. Two drawings with the same attractive components can implement contrary objectives.

The problem grows if the mechanism also has to reward earlier moves. Counters that record how many stones remain are of little help in reconstructing an earlier position after all the stones are gone. One can leave a trace beside each hollow, just as the human keeper did. A cam could sweep the hollows, act on only the traces that are set, and reset them when the operations are complete. Adding markers requires a supply. Taking them away requires a place for them to go. A rule retaining a minimum of one requires some way to stop at one. Carrying out one session without the keeper is a design exercise in energy, bookkeeping, and clearing up. None of it requires an unknown force. Much of it was invisible in the original list of materials.

The right to count an input as given

We will allow a person to present a starting position. Otherwise a question submitted to an electronic calculator would disqualify the calculator and a human opponent would disqualify a chess machine. The boundary is at the particular work claimed. If the apparatus is said to choose a move given the position, a person who records the opponent's move may supply a legitimate input. If the claim is that the apparatus can look at a board and recognize its position, a person who types in a perfect description has done the very work under investigation. Whether a hand is allowed depends on what the hand does.

Training raises the same issue over a longer interval. People can prepare a classifier by showing it labeled examples before turning it loose on unclassified cases. That can yield autonomous classification during

deployment. The historical account should say who supplied the training labels. If we claim in addition that the classifier improves in the field by finding out for itself when it is mistaken, the labels have to come from somewhere during that operation. An indicator that secretly summons a specialist to settle the difficult cases changes the claim. Calling the whole arrangement supervised or autonomous without identifying the period one is discussing can generate a surprisingly long argument between investigators whose observations agree.

Here is a harder example. A village healer occasionally looks over a weather learner's record and decides that all the forecasts made just before the waterwheel stopped should be excluded. The judgment might be entirely good. Yet if the machine's reported success improves substantially only after those decisions, its independent competence remains uncertain. Perhaps the builder can supply it with a mechanical flag for incomplete intervals in the future. Once that is done, future exclusions can belong to the device under an explicit rule. The earlier excluded trials still belong to a human analysis. Good accounting honors engineering improvements by specifying when they came into effect.

The boundary can even reverse the usual ranking of technological sophistication. A simple float in an out-of-the-way reservoir may genuinely perform all the sensing and corrective action claimed for it. A dazzling stage machine may depend on a hidden person for a crucial judgment while its considerable mechanical ingenuity goes into communicating that judgment to a spectator. The former has a smaller claim and firmer title to it. The latter may rely on more elaborate hardware simply to obscure who is making the decision. Credit bears little relation to the quantity of metal on show.

A practical way to audit a system is to follow information at a particularly inconvenient moment. Pick a situation in which two actions are possible and the correct one depends on something that has just happened. Write down the first physical difference between the world where one action is appropriate and the world where the other is. Trace that difference through the device until an actuator or an output diverges. If the trail vanishes into a person who is permitted to use specialist discretion, the person is part of that judgment. If it vanishes into a person who supplies an ordinary input that would be the same for any competing device, say so too. The method is less like assigning an intelligence quotient than following the provenance of a scientific specimen.

Testing two sorts of competence

There is a modest experiment that the seven-hollow machine makes particularly easy. Train one copy against an opponent who removes a single stone whenever possible. Train another against a player who has mastered the strategy of leaving multiples of three. Afterward let each play the other type of opponent. The operations can be simulated with completely specified rules, and one would report both training and test performance. High scores against the first opponent would have quite a different evidential value from consistent victories from winnable positions against the second. The general principle matters far beyond a subtraction game. Opportunities in the environment can sustain apparent skill as effectively as good internal organization can.

One can also evaluate improvement without predicting a particular learning curve. Inspect the two-stone hollow before and after many games. Have the machine play from two as a starting state. That separates knowledge of how to finish from the likelihood of arriving at two. A full-game success rate mixes both things. It is possible for the hollow to be well trained but rarely reached, giving a small effect on total wins. It is also possible to win often against a weak opponent while retaining a regrettable preference at two. Measures tied to individual decisions can reveal shortcomings hidden by spectacular aggregate scores.

Finally empty the hollows and rebuild their final populations from a written inventory. If play is statistically indistinguishable afterward, the transferable part of the learning is quite well localized. We could copy the preferences without repeating the games. If the experienced operator played better with the same inventory, we would have evidence that part of the acquired skill remained with the operator. A complete record of the marker populations is a surprisingly demanding test of our account of the machine. It distinguishes a working model of the cause from a colorful description of what was in the room.

Playing the comparison through

The seven-stone experiment is small enough to carry the test further than a thought experiment. For this book the exact update above was implemented in software, starting each legal move with one marker. The simulated learner

always went first from seven. Three kinds of opponent supplied very different educations. One always took one. One chose uniformly among legal moves at every turn. The third played the winning move whenever one existed and took one when facing a multiple of three, where perfect opposition leaves no forced victory. Training was repeated for 120 random seeds in each condition. After fifty or five hundred training games the marker populations were frozen for testing. This is a calculation about the stated rules, not an observation about actual stones or an attempt to reconstruct Michie's historical results.

The test itself can be exact for a particular inventory. With no tokens left for the machine to play, the opponent has just won, while a position of zero passed to the opponent means the machine has just won. Those endpoints have values zero and one from the machine's perspective. With one or more stones left, take the weighted average of the values of the positions to which a machine draw would lead. When it is the opponent's turn, use that opponent's rule to average or select its successors instead. Since every move reduces the total, the values at smaller piles are already available when computing the next. Apply this recursion through seven. Training was sampled, but a lucky run of test games contributes no noise to the reported win probability for each frozen inventory.

TABLE 1

What the seven-stone learner retained

Training opponent	Games	Takes one	Uniform	Optimal
Untrained	0	62.5%	51.6%	12.5%
Takes one	50	97.1%	57.2%	17.2%
Takes one	500	99.7%	59.6%	16.9%
Uniform	50	88.9%	84.5%	62.1%
Uniform	500	98.6%	97.7%	91.9%
Optimal	50	93.9%	82.6%	85.1%
Optimal	500	99.7%	87.3%	99.4%

Frozen win probabilities averaged over 120 training histories. Each column changes only the test opponent.

A machine trained for five hundred games against the always-one player reached almost certain success against that player. Against the specified optimal player it won only about one game in six. The same reinforcement rule with the same initial number of stones, trained against uniform moves, subsequently won about nine games in ten against the optimal player. Exposure to a variety of moves taught preferences that remained useful under a stronger form of opposition. More experience with the obliging opponent was hardly a substitute. Between fifty and five hundred training games the mean performance against the optimal player for that group even dipped slightly. One should avoid hanging a theory on so small a dip. The persistent gap across opponent types is plenty to explain.

The original untrained machine beats a perfect opponent only one eighth of the time. This number can be understood directly. Taking one from seven is the only winning opening against perfect play. The specified opponent takes one from six, leaving five. The machine must take two from five, leaving three. The opponent takes one again, and the machine must take two from two. Three equally likely correct decisions must occur, so the probability is one half cubed. Change the perfect player's arbitrary preference among moves from losing positions and it can give the learner a different curriculum while continuing to play optimally whenever a forced win exists. Calling an opponent perfect says less about the educational environment than one might suppose.

The result also reveals a limitation of the reported aggregate. A machine that nearly always wins from seven against the specified optimal opponent may have practiced one narrow route many times. The evaluator could begin it at four or five, or vary the opponent's choices at three and six, to examine how thoroughly it acquired the structure of the game. Against a uniformly moving opponent after five hundred games, the learner trained against optimal play averaged about 87 percent wins. That is good, but below the roughly 98 percent achieved there by the learner trained against uniform moves. The best education depends on the intended range of tests. “Stronger opponent” and “better teacher for every situation” are different propositions even for seven stones.

The simulation made a useful deliberate decision about how to report its results. It gave each training history equal weight, then averaged its exact test probabilities. Pooling only the successful training histories, or continuing until

a particular machine won a pleasing sequence and exhibiting that one, would inflate apparent performance. An inventor can legitimately exhibit an especially good artifact. A researcher who claims that a construction procedure *usually produces* effective artifacts must report the population from which the exhibit was selected. With small counters a few early draws can matter disproportionately. The distinction between the machine and the method of making machines is already experimentally accessible in a game that fits in one hand.

Notice the small reversal produced by doing this calculation at all. The original pebbles depended on a person to identify the outcome. In the software implementation removing the final stone triggered an update automatically. That program does meet the no-human-functioning condition for these games, even though its subject was a physical arrangement that failed it. A simulation can have its own genuine autonomous task without retroactively granting the simulated physical artifact a missing sensor. The machine whose workings we are identifying has to remain in view.

The fisherman could be absent because the arrangement of the trap allowed fish and tide to complete a capture. The water clock could be unattended because the level acted back on its supply. The learning game can retain a preference outside anybody's head, yet the proposed Stone Age version has to borrow an observer at the decisive moment. Its ingenuity has brought us closer to the subject of this book chiefly by making the remaining human work visible. One more lesson can be drawn without altering a single pebble. A good public demonstration of intelligent behavior should let the audience distinguish a hand that starts the trial from a hand that selects the answer. The next two cabinets made that distinction a matter of international curiosity.

CHAPTER 4

Two Chess Players in Cabinets

At the instant a chess piece begins to move across a board there is surprisingly little to distinguish two histories. In one, a person has seen the opposing move and found a reply. In the other, a collection of contacts has selected a circuit on the basis of the position. A linkage can convey either decision to the same arm. The rook arrives on the same square and confines the same king to the same edge. Knowledge of the game, however expert, cannot by itself tell an observer which history has occurred. The question about intelligence is a question about the path to the move.

Wolfgang von Kempelen's eighteenth-century chess attraction, commonly called the Turk, concealed a human player in its apparatus. Its imposing cabinet and mechanical figure were part of a successful piece of stagecraft. In 1912 Leonardo Torres Quevedo constructed an electromechanical chess player, El Ajedrecista. He demonstrated it publicly in Paris in 1914. It admitted a much smaller claims sheet than the earlier attraction. Given a specified family of positions, it played its king and rook against an opponent's lone king. Its apparatus determined how to respond. The narrowing of the game purchased a large gain in the independence of the machine. [7]

The work worth assigning in the first cabinet included some serious engineering. The mechanism allowed a concealed operator and a visible figure to act as one system. Keeping the source of the judgment obscure was part of the commission. Hand the operator an unfamiliar opening and the cabinet could benefit from whatever that player knew about it. Change players and the capability could change dramatically while the outward appearance remained familiar. Thus the machine is an especially apt control in an experiment about attribution. Removing the unseen player removes the source of chess judgment. It need not remove the cabinet's capacity for striking movements.

The second automaton makes the reverse thought experiment possible. Exchange its human opponent for a new opponent who plays quite differently. The machine's own procedure for responding stays in place. Remove the

opposition altogether and the public game stops, of course. That observation has sometimes led to muddled accounts of autonomy. A chess player, human or machine, is playing an interactive game. The other side's move contributes a problem at every turn. Providing the problem is a legitimate role outside the machine. Solving the problem on the machine's behalf is the contribution that would change the interpretation.

How much chess was in the cabinet?

The original version used a vertical board with pegs for pieces, moved on the machine's side by a mechanical arm. The starting conditions placed restrictions on where the three pieces could go. The opposing king had some freedom of initial placement, and considerable freedom of subsequent movement. A later version, often illustrated in histories of computing, used a horizontal board with electromagnetic movement from underneath. In that version a recorded voice could announce check or mate. A picture of the later machine beside a date belonging to the earlier one can give a reader the wrong artifact to imagine. We are interested primarily in the continuity between them: a physical move by the opponent produced a different reply through a rule embodied in the device. [7]

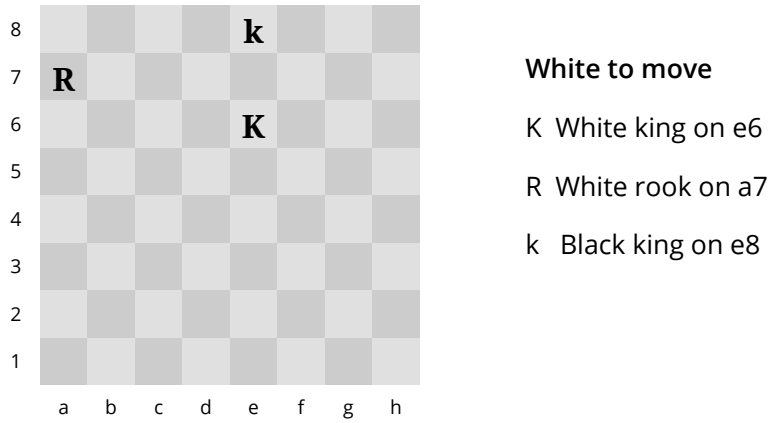
A king and rook against a bare king leaves enough of the game intact to make its scope intelligible. Kings move one square at a time and cannot end up adjacent. A rook reaches along a rank or file. Its ability to command a whole line can restrict where the opposing king may go, but it also needs to be kept from capture. The stronger king can help protect the rook and push the opponent toward the edge. There is still a delicate distinction between completing mate and merely leaving the opponent without a legal move when the king is not in check. That last condition produces stalemate. Naming these constraints does more to describe the machine's achievement than calling it a chess genius.

Here is a small illustrative position, independently of the sequence implemented in the historical cabinet. Place the black king on e8, the white king on e6, and the white rook on a7. Let White move. A rook move to a8 checks along the eighth rank. The white king on e6 covers d7, e7, and f7. A black king trying d8 or f8 would still be traversing the rook's line. Black has no piece to interpose and cannot take the distant rook. It is mate. The conclusion can be checked square by square with a diagram drawn on the back of an

envelope. Now move the white king farther away before carrying out the same rook move. A square on the seventh rank may be available. The same local action by the rook no longer does the same job.

FIGURE 1

One rook move, three pieces



Illustrative position before White plays Ra8. The white king guards the three seventh-rank escape squares.

What would a machine have to know to produce the first outcome? One possibility is an exhaustive list of permitted positions and moves. Another is a small body of rules that divide the board into regions according to relations such as whether the kings oppose one another and whether the rook is threatened. A third is to look ahead by trying moves in an internal representation and testing the consequences. All could produce the identical mating move in our illustration. Only one of them might be practical in the chosen hardware. We should not infer a specific internal process from the beauty of the final move. Torres Quevedo's machine was built around a restricted algorithm. A winning reply is good evidence that some relevant procedure has been executed. A series of controlled changes to the position would tell us much more about that procedure than a single photograph of checkmate.

The rules of a demonstration also deserve inspection. Historical accounts indicate that the procedure could take as many as 63 moves, and the conventional fifty-move draw condition complicates any claim of guaranteed tournament success. The admissible starting positions themselves were restricted. This does not deprive the machine of its achievement. It gives an

engineer a meaningful contract to evaluate. A machine that wins a restricted endgame by a fixed procedure is a different proposition from a machine that can play an ordinary game from the initial position. In the history of intelligent devices, a missing subordinate clause can enlarge an achievement by orders of magnitude. An honest account puts the clause back. [7]

An oracle with an instruction manual

Several centuries before either chess cabinet, a North African scholar considered the capacity of a device to answer a question. Ibn Khaldūn discussed the zā’irja, a divinatory procedure involving letters, numerical relationships, and a diagram. Accounts of a demonstration in 1370 or 1371 describe a teacher suggesting that the device itself be consulted about the antiquity of the procedure. The reported answer, given in verse, associated the practice with Idrīs. The historical evidence concerns a performance and its interpretation, including an operator versed in the rules. A verse naming an ancient figure establishes neither the date of invention nor the reliability of the instrument as an independent historical witness. [29]

The episode is a particularly instructive challenge to the criterion of this book. There was a question in ordinary language, a material aid, transformations governed by a tradition, and a reply in words. In that superficial respect it resembles a person addressing a conversational model. Yet if a learned operator had to perform the transformations on each occasion and determine what the resulting text said to the questioner, much of the answering lay in the trained practice. The artifact could hold a relation stable or guide the ordering of operations without autonomously completing them. Removing the operator from the historical demonstration and observing what still generates a verse would be a more exact test than deciding which modern device the arrangement most resembles.

It would be just as rash to infer that the operator simply invented an answer at will. A rule followed by a person can severely constrain the outcome. Lines on a chart, conventional correspondences, and the letters of the question can jointly determine operations that a particular individual would never have devised. A tradition can give a question surprising consequences even when hands and judgment are needed to draw them out. A historically serious account can distinguish a partially specified algorithm, the person's execution of it, the interpretive latitude remaining at particular

stages, and the religious or intellectual meaning attached to the performance. No single comparison to a chatbot encompasses all of those things.

The question about antiquity has an especially modern resonance. Ask a contemporary system what information it was trained on or why its response took its particular form. Even a genuinely independent reply can lack access to the facts required for an authoritative answer. An instrument can faithfully perform the operation it was designed to perform and return words about its origins that have no independent historical warrant. Autonomy of output and privilege concerning one's own construction are separate attributes. A scholar evaluating the zā’irja's response could well insist on an external record for a claim about its age. An investigator evaluating software can insist on design or training documentation for a claim of equivalent scope.

The zā’irja, the Turk, and El Ajedrecista thus represent three opportunities for misclassification. In a documented interpretive practice, dependence on a trained human may be overt and legitimate. In a stage illusion the position of the human chooser may be deliberately obscured. In a restricted but genuine automaton the human opponent is plainly visible while the choices assigned to the mechanism really do take place within it. The presence of a person in the room is a poor discriminator among the three. What matters is the causal job the person is performing and the claim being advanced about the resulting answer. That is why our criterion can admit a prompted language model while excluding a tally kept by hand, and why its application sometimes requires surprisingly careful historical reading.

The choice the designer left open

Consider the objection that every reply was already in the inventor's head. Even when a designer can prove that a method will work for every admissible input, the next concrete action need not be known at construction. A loom's pattern can specify in advance what should happen on row thirty-one. The chess procedure can specify what to do *if* the opponent reaches one situation and something different if the opponent reaches another. To predict the exact movement of the arm from the construction documents alone, one would still need the opponent's choice. Conditions selected after construction enter causally into the action. This is a small opening through which an external world can enter a mechanism made of fixed parts.

The word choice here need not carry a theory of free will. It identifies a functional relation. There are at least two physically distinguishable inputs, two allowable responses, and an organized dependence of the response on the input that serves an assigned purpose. An electrical selector can implement that relation. Whether a human experiences something richer in moving the rook is an important but separate question. Insisting that an apparatus experience indecision before it can have selected anything would make ordinary descriptions of railway interlocking, message switching, and automatic sorting difficult to understand. Equally, using the same verb does not give the cabinet a private experience of triumph. The game gives the engineer an unusually definite criterion of competent selection.

The engineering of the input deserves as much admiration as the engineering of the rule. In human chess a piece can fall on its side or sit partly over a line. People often adjust it without debate. Wiring and contacts require a commitment about what counts as a legal position and when a move has ended. If the machine samples while an opponent is lifting a king from one square and has not yet set it on another, it can momentarily inhabit a world without that king. A sensible interface must impose an order on events or make the signals mechanically unambiguous. This general consideration need not be a claim about the details of a surviving Torres circuit. It explains why turning an elegant algorithm into a machine that will operate on a public board is its own invention.

One can see a related difference in the roles of enforcement and strategy. An indicator that warns an opponent of an illegal move performs a small recognition task. A mechanism that responds to a legal move performs a strategic task. An actuator that moves the white piece completes the physical one. They need not be independent pieces of equipment, but they are independently testable claims. A prototype could choose a sound move and fail to move the rook because the magnetic coupling was weak. Another could move every piece smoothly while quietly admitting illegal positions. A photograph of the machine at rest reveals almost nothing about how these failures would distribute themselves.

A game that knows when to stop

The possibility of a terminal condition is more important than it may look. A cabinet that always moves the rook when the other king moves might continue making legal-looking actions forever. To complete a game it must in some manner detect a condition under which play is over. Someone who stops a display only after seeing a pleasing position may be acting as a stage manager or as the only functioning terminal detector. To allocate credit we must know which. In the later version of El Ajedrecista even the theatrical announcement was attached to an arrangement built for the game. Mechanical drama and machine competence were capable of sharing a cabinet without being identical.

Chess is attractive to experimenters because several difficult engineering matters have already been solved by the rules of the game. Board geometry divides the world into a finite set of places. Players agree on which changes count as moves. A king that is mated can be identified from the position and the rules. In the laboratory of a robot scientist the analogues of those agreements can become the largest part of the labor. Is a slightly faster rate of growth a positive outcome? Can the same observation be produced by a faulty instrument? Should an ambiguous sample be excluded and, if so, by whom? The clarity of chess allows investigators to isolate certain capabilities while leaving the machine invitingly silent on many questions faced outside a board.

One could test how the restricted chess player's competence travels beyond its training ground. Replace the symbols by colored blocks while leaving the contacts and allowed motions unchanged. If operation is preserved, the mechanism's dependence is on encoded positions, not on a visual conception of royalty. Exchange the rook's rules for a bishop's but leave the mechanism unchanged. Correct mechanical responses will now routinely make errors in the new game. Add an extra opponent's piece that can block a line and much of the earlier proof ceases to apply. The machine was built for a world with a certain structure. The fact that its behavior there is intelligent in a narrow sense places no automatic claim on transfer to a world whose vocabulary has changed.

Modern demonstrations can be subjected to similarly plain changes. If a system that answers questions about a diagram continues giving the same answer when a crucial arrow is reversed, perhaps it relied on a textual cue. If a

question-answering system performs beautifully on familiar chess puzzles and fails when files are renamed, it may have learned a useful association without acquiring a portable procedure. Neither test needs to establish a universal theory of intelligence. A good test tells us which particulars in the display actually determined the answer. At the chessboard the point is almost literal. Lift the wrong peg and watch what the mechanism does.

The transition from the Turk to Torres Quevedo is not a simple story of deception yielding to scientific truth. A narrower and more fruitful lesson concerns the way we describe a whole performance. The humans who engineered a mechanism deserve credit for the accuracy of its mechanical display even if a hidden expert supplied the game. The inventor who encoded an endgame method deserves credit for an algorithm that the mechanism executes even when the opponent supplies surprise. A later critic who points out the starting restrictions can make the historical achievement more intelligible. Tasks have internal divisions, and the divisions matter most just where the public sees one seamless motion.

A final comparison looks ahead to a learned player. Give the Torres apparatus its hundredth opponent and its method is still the one put in at construction. Any improvement from adjustment or repair belongs to whoever made it. Give a program a way to change its evaluation of a board according to the subsequent course of play and its hundredth opponent might encounter a different judge. That is one way an experience can become part of a future decision. Ross took a smaller physical step toward this possibility. Instead of asking which move to make on a crowded board, his machine could answer a much narrower question about a place where its own tip had already been.

CHAPTER 5

When the Maze Writes the Memory

In the drawings of Ross's machine, the arm has the proportions of a little derrick. Its freedom of movement is modest. One pivot allows the arm to rise and fall. Another allows it to swing sideways. The potentially informative part of this landscape lies in four passages on a vertical plate. A tip can enter one, meet the sloping surface at its end, withdraw, and move to another level. It is tempting to describe this as a mechanical rat running through a maze. The rat then occupies all the mental space and the experiment disappears. A probe on an arm that samples four distinctive passages is a better instrument for studying the origin of a particular change in behavior. Its simplicity makes it possible to ask exactly where the change came from. [1]

Take the lowest passage in the account. The arm starts at the left. Applying power moves it to the right. Its position brings one memory cell into communication with the moving assembly. At the far end the deflection of the tongue makes an electrical contact. One circuit initiates withdrawal. Another operation leaves the relevant cell conditioned. When the arm comes opposite that passage on a subsequent run, the contact associated with its position can initiate a withdrawal earlier. It no longer needs the distant surface to supply the signal at the old point. A stimulus encountered late on the first traverse has enabled an earlier stimulus to govern a movement on the next.

There is a distinction here between knowing with words and being ready to act. No sentence inside the machine says that the lowest slot is blind. The readiness to withdraw encodes one practical aspect of the slot. An engineer could read the switch setting and write a sentence about it, just as a traveler might look at a bent road sign and translate its arrow into directions. The machine's immediate use of the setting does not require that translation. Calling the setting a belief may introduce more philosophical machinery than four passages warrant. Calling it retained information is already a substantive

claim. It depends on distinguishing which of several encounters has occurred and preserving a difference that will govern later motion.

The experiment that begins at the start again

Starting a second trial is an experimentally productive thing to do. On an ordinary first passage we cannot see whether the machine has retained anything. Even a dead end can look like intelligent avoidance if we only see the arm after it has already hit the obstruction. On a second passage the important event takes place earlier. A previous encounter has become causally available at a time when the original sensory stimulus is absent. A video shown in reverse order would seriously misdescribe the result. So would a photograph assembled from a first-run start and a second-run finish. History has become one of the inputs to the response.

There are several ways to make that claim stronger in a contemporary reconstruction. Record the exact maximum penetration into each slot in each trial. Change one cell's state while leaving the position of the arm and the other cells unchanged. Exchange the state of two cells and ask whether the respective responses follow them. Alter a sloping surface between trials. A machine with a stale memory may start to withdraw from a newly open passage. The error is evidence for the proposed account insofar as its location and timing were predicted. We often learn more about the location of memory from a systematic mistake than from a smooth run in surroundings that have remained cooperative.

Each proposed intervention has a nuisance attached. Changing the cell can jostle the arm. A different surface can introduce new drag. An observer estimating depth from a slanted camera view can make a fixed movement look shorter. Good engineering experiments try to arrange harmless manipulations that reproduce the nuisance while preserving the variable of interest. In a reconstruction one might touch a neighbor switch without changing its state or allow the machine to cool for the same interval before every trial. The difficulty is less formidable than it sounds. It is the ordinary price of claiming that an electrical state, rather than friction or an operator's hand, explains the changed path. Ross's publication gives us the rare advantage of knowing which controls his circuit suggests.

His first and second plotted paths deserve to be kept separate from his third drawing. The refinement he discussed could suppress excursions into blind

passages altogether. The device he described still made small entries. The difference is only a short stroke on the page. It is the difference between a demonstration and a design. Early technological history can develop an unfortunate perfection in retelling: an engineer explains a nearly achieved improvement, an intermediary illustrates the improved configuration, and later readers inherit an ideal machine with a convincing publication date. Ross's precision gives us the means to resist that process and to enjoy the achievement he actually reported. [1]

The four memory cells should not be asked to remember arbitrary journeys. A fixed arrangement links a cell to the apparatus's position when a passage is approached. That supplies much of what a nervous traveler would have to learn in a strange landscape. The probe need never infer which corridor is which from scraps of sensory evidence. Position is effectively encoded by the construction. If the entire plate slides while its contacts remain fixed, the remembered lesson may be recalled at the wrong place. This is an instance of a general problem in navigation. A remembered direction is useful only if a system can tell when it is at the location to which the direction refers.

A mouse with memory beneath its feet

By about 1950 Claude Shannon had built Theseus, a small maze-solving apparatus in which the visible mouse was guided by machinery associated with the board. It explored a maze with movable partitions. Shannon's published presentation explains the storage in unusual detail. Each square had a direction memory represented by two relays. Two binary settings can distinguish four possibilities, a comfortable match for north, east, south, and west. The machine kept the direction in which the sensing finger had last left each visited square. Different rules governed exploration and the later attempt to use the stored directions to reach the goal. [8]

The phrase “last left” deserves attention. One could instead store the first direction ever tried, the direction most frequently taken, or the direction pointing toward the apparent geometric center of the goal. Those rules would give quite different behavior. Suppose a traveler leaves a junction to explore a dead end and later returns to that junction. Upon finally departing along a more useful branch, the last-departure record is overwritten. It is the overwriting, something that can sound like a failure of memory, that allows a

finished path to be read from the record. There is a compact piece of mathematics inside this engineering decision.

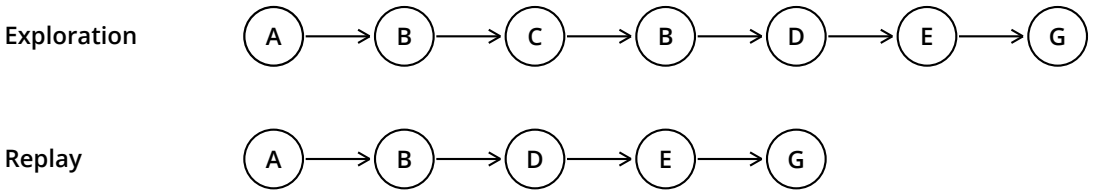
Imagine labeling departures by their time during an exploratory walk that eventually reaches a goal. At the starting square follow the direction of its last departure. That takes you to another square reached after that departure. Unless it is the goal, its last departure must be later still. Continuing the procedure yields a strictly increasing sequence of departure times, all from the finite completed walk. They cannot form a cycle. Eventually one reaches the goal. The apparatus need not store a clock reading at each square. The ordering is a property of the way the memory was written. A surprisingly thin record can preserve a route because the laws of the update and the history of the walk exclude alternatives.

This argument applies to an unchanged maze and the visited succession for which the last exits were recorded up to reaching the goal. It does not imply that the recovered path is the shortest possible path. A different exploration could have taken a shorter corridor that this walk never saw. Nor does it give directions from a square the apparatus never visited. What is guaranteed by the argument is much more interesting than a general compliment about remembering. A local record that appears to carry no global route can, under stated conditions, encode a continuous way home. The route is distributed across many locations. No individual relay has to know the end of the journey.

Consider a tiny written example. Let an exploratory path visit A, B, C, B, D, E, and end at G. The last exits recorded are A to B, B to D, C to B, D to E, and E to G. Starting again at A gives A, B, D, E, G. Starting at C after the run gives C, B, D, E, G, a route that emerges from the same little store of decisions. A record of the whole list of movements would also work, but replaying it from a new starting point would require finding the correct place to enter the list. The local representation solves a small indexing problem. It spends storage at every square to make the next direction immediately available wherever the mouse is placed.

FIGURE 2

Recovering a route from last exits



At B the final departure toward D overwrites the earlier departure toward C. The figure illustrates the memory rule using an invented walk.

There is a cost to this cleverness. Suppose two identical-looking junctions actually lead to quite different destinations. If the machine identifies its position only from their appearance, it can retrieve the wrong memory. Theseus enjoyed a fixed association between a square and the circuitry that served it. The experimentalist had supplied a grid with an addressing scheme. If the historical claim is that the apparatus found a route in its given maze, that is perfectly legitimate. If one imagines it identifying its own location on a trackless moor, the most difficult capability in the new story may be the one nobody had to solve on the board. The scope of a demonstration is written as much in the floor as in the moving vehicle.

The old route can do harm

Shannon showed what could happen when he moved partitions after the machine had learned a solution. A familiar sequence of squares now led back to its beginning. The apparatus could follow its stored preferences around the loop. In the reported demonstration a counter interrupted this behavior at a count of twenty-four and the machine returned to exploration. Members of the audience proposed psychological analogies. The circuit had access to excessive duration, which was enough to provoke recovery. The apparatus did not have to formulate an explanation of its predicament in order to escape it. [8]

It is worth sketching what breaks in our little proof when the maze changes. The temporal argument assumed that following A's remembered direction now delivers the traveler to the same next square that followed its last departure during training. A new obstruction can make that step unavailable.

A detour or a rule for handling the obstruction introduces a fresh departure that is absent from the old chronological sequence. The strict increase of historical times no longer applies. Keeping every remembered bit unaltered can preserve a completely accurate record of an obsolete world. A reliable memory can contribute to unreliable action when the relation between record and surroundings has changed.

Counting movements is a canny but blunt way of discovering trouble. If a legitimate new route is lengthy, the same limit could interrupt good behavior. If a loop is very short, several useless revolutions may precede the interruption. A more elaborate device might keep track of repeated visits to the same square with the same intended direction. That requires additional memory and a rule for clearing it. Another might test a suspect wall once, amend only the corresponding local record, and preserve useful directions elsewhere. A third might retain old and new maps and choose between them on evidence of recurring configurations. These are proposed improvements to a class of machine. The historical source distinguishes what Shannon's relays did from the tempting additional functions an onlooker can imagine. [8]

The trade becomes visible in a house where a familiar hallway is occasionally closed. Forgetting the entire house on meeting a locked door wastes a great deal of prior experience. Following the original route for twenty-four steps before admitting there is a problem can be costly as well. Retaining both arrangements pays if the closure is intermittent. The benefit of preserving the old route depends on the rate at which the world changes back. Learning and forgetting are solutions to a particular distribution of environments. Without describing that distribution, there is no single optimum amount of fidelity to experience.

A maze that already contains its solver

Consider a photograph showing one living tube connecting two pieces of food through a labyrinth. It is tempting to imagine a little organism setting out from the entrance, taking wrong turns, and finally arriving at lunch. That is the wrong starting photograph. In the experiment reported by Toshiyuki Nakagaki, Hiroyasu Yamada, and Ágota Tóth in 2000, a slime mould, *Physarum polycephalum*, already occupied the maze when the investigators supplied food at two endpoints. The organism could withdraw material from passages

and maintain connections elsewhere. What appeared in the later photograph was the route that survived this reorganization. [32]

There was an instructive difference between two choices within the same apparatus. At one branching section the alternative routes differed substantially in length, by about 22 percent relative to the longer route. The shorter connection was strongly favored. At another section the lengths differed by only about two percent. Variations in the course taken by the tube were large enough to obscure that difference. The investigators' table also records instances with multiple connections or no selected route. Look at the table as well as the successful photograph and the claim becomes more interesting. The material that solves the problem sets a practical limit on which differences in path length it can resolve. [32]

Shannon's mouse could travel down a passage and change a relay setting. One could interrupt its journey at a junction and ask which setting would send it home. The plasmodium confronted alternatives distributed through its own body. Removing a portion of it would change the very system that maintained the alternative passages. In both cases the investigator constructed a labyrinth and fixed what was to count as arrival. Neither investigator had to stand at the critical fork pointing out the shorter turn. Yet the physical meanings of “explore,” “choose,” and “remember a path” are quite different in the two experiments. The final geometry alone conceals those differences.

Does the living arrangement belong in a history of artificial intelligence? Under the builder's rule it makes an exacting boundary case. Humans built the apparatus and arranged the task. They did not invent the organism's capacity to move its own material around. If the object of study is the whole experimental device, its construction was an act of engineering. If the object of study is the origin of the capacity that selected a route, much of the explanation lies in biology. We can draw either boundary and then check what falls within it. Drawing one for the artifact's ancestry and silently switching to the other when assigning credit would give the engineers a remarkable achievement that they never claimed.

A further test would ask what endures when the task changes. Seeing a thick tube along a short route establishes a spatial result in one maze. To establish transferable learning, one would have to specify what carries over to another trial and compare performance with a suitable fresh start. A reward for fast arrival, a reward for keeping two alternate paths open, and a reward

for the least total living material would make different demands. Even in a maze with only two food sites, the person who writes the scoring rule has more latitude than the word “solution” usually admits.

Good tests for a machine that remembers

One can measure the benefit of learning a maze in more than one currency. Count the moves to first discovery, the moves on a second attempt, the power consumed while exploring, and the time lost after an unannounced change. A design that spends considerable effort on its first survey can do well on every subsequent trip. A design with small memories and quick local decisions might be preferable for a maze that will last only two trips. The design that looks brilliant on the second journey might lose to a simpler competitor if one includes the original cost of setting up the memory grid. Which comparison matters depends on the service the machine is expected to provide.

A more revealing demonstration would draw two new starting squares after training. Pick one visited square and another never reached by the exploratory walk. A machine with a stored response at the first may behave quite differently at the second. Give it a symmetrical maze rotated through a right angle without corresponding rotation of its direction memories. A human familiar with the geometry might transfer the solution. A machine whose representation is tied to absolute directions may have to relearn. Transfer across starting places and transfer across rotations are different achievements even if both are advertised as flexibility. Each calls for a precise description of what changed and what remained available.

We should be equally precise about an apparent success in a changed maze. Suppose a moved partition opens a shortcut. The old route still reaches the goal. A machine that remains faithful to it has suffered no obvious failure. Its twenty-four-step precaution may never be invoked. Yet it may remain ignorant of a large improvement. Error-triggered exploration finds some changes very well and overlooks others completely. The environment can become better without announcing that one's current decisions are wrong. The corresponding difficulty for a laboratory robot is easy to recognize. It may continue conducting decisive experiments among the hypotheses it already knows while never considering a still better explanation outside the catalog.

There is a fruitful comparison between the maze memory and the ant trails that form in natural foraging. In certain ants, successful foragers add

pheromone to a route and unreplenished trails fade. A newcomer can then benefit from journeys it never made. In the engineered maze a local memory points in a direction. In the forest an odor can change the probability of taking a path. Tracing either mechanism establishes a contribution outside the body of a single traveler. It leaves open the important differences between a relay that keeps its latest setting and a chemical trail whose strength reflects contributions and decay over time. The artificial device remains our subject, and the comparison helps us locate its special economy of memory. [9]

Now remove the little mouse and leave the circuits in the table. Have we extracted the intelligent part? Only partly. Without the means to sense a wall and move, the table can retain route information but cannot acquire more by exploring. Put the mouse on a plain board and it has lost the very local records on which replay relies. The experimental unit is an arrangement whose moving member, grid, switches, motors, and external partitions cooperate to produce the behavior. It is possible to draw an engineering boundary around such an arrangement without pretending that the most photogenic component contains all its faculties.

The contrast with Ross is particularly illuminating. Ross's probe could become conditioned at one of a small set of prearranged approaches. Shannon's apparatus could use records attached to many places to assemble another journey as it moved. One could call each device a learner and be right in a broad sense. The name alone would tell a future builder very little about which components to preserve. The first question for a reconstruction is whether a prior encounter can govern an earlier reaction. The second is whether local last-exit records can be read as a route and under what disturbances that guarantee fails. Intelligence becomes an engineering subject when it yields questions of this specificity.

The next apparatus had still another relationship to a changed world. Ashby's homeostat was built to vary its own internal couplings in search of stability. It might recover from a disturbance without learning where a wall stood, what direction it had traveled, or which switch any previous configuration had required. It offers an opportunity to ask how a device recognizes an acceptable internal condition before crediting it with an account of the world outside it.

CHAPTER 6

A Machine Searches for Its Balance

A visitor who wants to see a great deal of activity from Ashby's homeostat should arrive when something is going wrong. When it finds a workable organization, the part that searches for a new one can come to rest. The quiet of a satisfactory run makes an awkward demonstration. A spectator inclined to identify intelligence with continuous novelty might conclude that the investigator has succeeded just when the machine is doing least of interest. Many useful systems are paid, in effect, to reach this condition. The household would much rather have a refrigerator whose contents stay cool than one that continually announces ingenious new approaches to refrigeration.

W. Ross Ashby investigated how an organism or a machine might preserve certain essential variables within acceptable bounds. His homeostat, built in the 1940s and described in Design for a Brain, had four interconnected units. The state of one affected the others. In the configuration he describes, stepping switches could select among settings of parameters governing the interactions. If a variable moved beyond limits, a change of setting could alter the arrangement responsible for its behavior. Ashby derived the available parameter settings from a table of random numbers during construction. We should distinguish that historical fact from an assertion that a mysterious source of chance invented a fresh arrangement at every step. The available alternatives and the machinery for selecting them had a builder. [10]

Two ways of changing a response

A simple regulator already changes its response. Raise an engine's speed and its governor changes the amount of steam admitted. The relationship between speed and corrective action can remain much the same throughout its life. Ashby wanted to examine the consequences of adding a second way of changing. The relations responsible for ongoing behavior could themselves be altered when that behavior left the system in an unacceptable condition. A

particular connection might exert a larger influence, a smaller influence, or an influence with a different sign. Following such a change the same disturbance could play out in a system with different dynamics.

An everyday illustration is a shower in a house with temperamental plumbing. A person turns a handle a little when the water is too cold, waits, and turns it back if it grows too hot. That is regulation. After several uncomfortable swings the person may discover that the supply takes a long time to respond and begin allowing a longer interval between adjustments. A different feature of the response has changed. Suppose the surprise is more fundamental and the hot and cold connections have been reversed since yesterday. Making larger adjustments in the old direction can worsen the problem. Changing the sign of an adjustment may restore control. Our example contains a human operator for ease of imagination. In Ashby's apparatus the question is which equivalents of such changes a device can make by itself.

Consider a deliberately simple mathematical world that helps explain the search. A deviation of one unit at one instant produces a deviation of magnitude one half at the next under configuration A. Repetition gives one quarter, one eighth, and ever smaller deviations. Under configuration B a deviation of one produces a positive deviation of one and a half, then two and a quarter. One operating rule tends to suppress a disturbance. The other amplifies it. A controller capable only of making a minor correction within B can remain in trouble. A second mechanism that changes B to A when the deviation passes a limit can find a workable system without identifying why B was so poor. The numbers in the example describe an invented recurrence. They help isolate the logical achievement of a device that can change its own couplings.

Stability is a property of an arrangement together with an environment and a time scale. Configuration A may look excellent when the incoming disturbance is one unit at a time and terrible when a persistent displacement drives it faster than it can respond. A physical needle has inertia. A damper can turn an oscillatory approach to a value into a slow smooth one. Excessive delay can cause a previously harmless response to overshoot. When the homeostat finds a configuration that keeps its variables in range, the very circumstances that made it acceptable belong in the report. There is no universal list of settings called good independent of what the system will face.

A criterion for saying “enough”

The clever part of the homeostat is how little it needs to know about the reason a setting works. It requires some way to distinguish an essential variable inside an interval from one outside it. It requires a way to alter the arrangement. It can then stop altering while successful operation continues. A mechanism that merely produces random configurations forever has two of these capacities at best. Without retention conditional on the result, a successful configuration will disappear as arbitrarily as an unsuccessful one. A mechanism that always keeps its original setting, on the other hand, can succeed only through prior design or good luck. The selective stopping supplies an important asymmetry.

For the four switches with twenty-five positions each Ashby gives 390,625 possible joint settings. That count follows from multiplying twenty-five by itself four times. The count of arrangements alone cannot predict the time to success. If half of the possible arrangements are acceptable in the present environment, searching could be easy even with many settings. If one arrangement in a thousand is acceptable, it could be difficult with much smaller hardware. The way the switches advance matters too. A procedure that revisits the same unhelpful sequence can do worse than one that retains a record of places searched. The historical homeostat must be judged by its actual mechanism. Multiplying its choices is a useful inventory and a decidedly inadequate measure of its intelligence. [10]

Here is an intentionally different search rule whose cost we can calculate. Suppose each new setting is an independent uniform draw from a collection in which one setting out of sixteen works. The chance of missing every acceptable setting in ten trials is fifteen sixteenths raised to the tenth power, approximately 0.52. After thirty-two independent trials it is still about 0.13. The average number of trials to first success is sixteen. Those statements are compatible. An audience watching ten attempts has a substantial chance of seeing no success even when the machine is working exactly as designed. The calculation would give poor guidance if the settings are strongly correlated or if the environment changes before each trial. It is a model of why demonstrations of search need a specified time allowance.

Observe the importance of a cost the demonstration scarcely mentions. During each attempt, essential variables may be outside their acceptable

ranges. A machine whose bearings tolerate five minutes of bad vibration can explore more freely than one whose bearings fail after three seconds. A software game player able to start a fresh game after a loss can afford great liberties. A spacecraft may be permitted no comparable losing game. In a living organism many variables have real limits as well. A search that succeeds with high probability after enough trials may be physiologically or mechanically useless if a common early trial destroys the searcher. The homeostat brings a question about uninterrupted survival into a field often preoccupied with eventual success.

A second limit is the supply of workable configurations. Suppose the sensor is disconnected, or every available setting produces positive feedback in the altered environment. Further switching can spend considerable energy without ever reaching a useful region. The absence of a satisfactory answer among the possibilities is different from taking a long time to find an answer that exists. An observer needs some independent way of distinguishing the two. Looking through the possible couplings may help the builder. Nothing about repeated motion of the selector by itself establishes that the selector is approaching an answer.

Who supplied the purpose?

The essential variables in Ashby's construction are essential because someone designed the machine around them. That deliberate choice can appear to disqualify the example to a philosopher who wants a device to supply all its own purposes. It is entirely compatible with the rule being investigated here. The builder may state the objective. What remains to be established is who determines which arrangement meets that objective in the situation actually encountered. A machine that can make the latter discovery has done something that a preadjusted fixed controller may fail to do.

We can see the difference with a tale of two repairers. Each is given a device with a row of switches and a meter marked with an acceptable band. The first consults a chart supplied by a designer and selects the setting for the stated load. The second tries available settings until the meter stays in the band. Both may end with the same switch positions. If we exchange the loads without changing their labels, the first repairer's competence depends on the chart remaining applicable. The second procedure may discover another working setting. If we replace the repairers with machines that execute their respective

procedures, we can still assign these different capabilities. A goal inherited from a designer is consistent with some independence in discovering a means.

The observer's difficulty lies in distinguishing valuable stability from a convenient surrogate. A meter can remain in the middle because the plant is healthy. It can also remain there because the wire to the meter broke at a convenient moment. A machine that optimizes its registered condition can be exceptionally satisfied by that second event. The word satisfied is shorthand for a halt in the search. Whether the real condition justifying the device's existence has been met is a question about the relation between measurement and world. This is why fault diagnosis will demand something beyond keeping a reading in a preferred interval.

Suppose a modified homeostat has a battery meter as well as a position meter. It can find regimes where its position stays in range by running a motor continually and using up its battery. An observer who admires an hour of perfect regulation may be watching the systematic destruction of the possibility of a second hour. A fuller device could treat battery state as another essential variable. Then there might be disturbances for which preserving both position and battery is impossible. Choosing which requirement can give way has an ethical analogue in human institutions but here is already a concrete engineering problem. Someone has to authorize a trade. Adding another gauge brings another criterion into the apparatus. It need not resolve conflicts between criteria.

This is a reason to resist treating the sequence of machines in this history as a ladder with humans inevitably at the top of each rung. One machine inherits an excellent criterion and a meager method of search. Another inherits a richly detailed criterion that is actually a proxy of questionable relevance. A third has a few physical conditions on which its continued operation genuinely depends. Their degrees of autonomy, power of representation, and relevance to the world vary independently. If we measure all three by the number of different responses they can make, we lose the ability to distinguish a self-preserving system from a random switchboard.

A workable configuration is a kind of memory

Leave a successful homeostat running and ask what it remembers. It can retain the switch settings that currently work. It need not be able to give the entire history of the configurations that failed. Change the environment and it

may resume searching. Return the environment to its original condition after an intervening search. Whether the original arrangement can be recovered promptly is a new empirical question. A machine that keeps only its currently successful setting has a very different temporal capacity from one that stores a portfolio indexed by recognized environments.

Imagine a workshop with two recurrent loads, heavy in the morning and light in the evening. A homeostat-like mechanism could spend the first portion of each session finding a usable arrangement and operate well thereafter. A device fitted with a reliable indication of which load is connected and two independent memories could recall a successful configuration immediately. On the other hand, the extra classification can do harm if the supposed heavy load and light load no longer predict the relevant dynamics. The portfolio gives faster transfer across familiar regimes at the cost of having to identify a regime correctly. In a small workshop it may be worthwhile. In a continually changing environment the older solution might turn out to be more robust.

The example is also a way to distinguish learning about a world from fitting oneself to it. A thermostat that has found a setting need not carry a temperature model of the house. Similarly, a homeostat that keeps its variables in range can achieve an effective relation to its circumstances without constructing a map of those circumstances. A map would be valuable for explaining an unseen change, trying alternatives without disturbing the apparatus, or predicting what will happen to a different variable. It costs something to maintain, and in a stable system it can become an expensive ornament. The historical machine draws attention to models partly by showing how much a device can do with so little representation.

There is room here for a more exact use of the word intelligence. We might reserve it for systems that can construct and test an internal alternative before acting. That would exclude much of the homeostat's behavior but admit some chess algorithms and much laboratory planning. Alternatively we might include any mechanism that can search its own possible organizations and retain one on the basis of its consequences. Then the homeostat is a clear candidate. Our historical argument does not require the reader to choose once for all. What would be intellectually costly is to attribute internal planning merely because a trial-and-retention process has found a good configuration, or to deny the reality of its discovery merely because a human supplied the meter.

Before turning away from the quiet apparatus, it is worth attempting one last diagnostic maneuver in imagination. Place an identical-looking unit beside it with the successful switches glued into position. Provided the conditions stay favorable, the two might behave indistinguishably for a long time. Change the environment modestly. Both may continue in bounds. Change it in a way that makes the old configuration unstable but leaves a workable alternative available. The glued machine has lost a resource that was invisible in its behavior during the entire preceding demonstration. Independent capacity is often best defined by the ability to survive a particular class of changes. Part of the historian's work is finding out whether anyone actually imposed one.

The route from these switching devices to learned evaluations is shorter than it first appears. Rather than asking whether a whole arrangement can keep a needle in range, a builder can provide a collection of adjustable influences whose sizes alter the machine's judgment. A feedback signal might then alter just those influences implicated in an error. The new arrangement could influence many future cases, including ones never encountered during adjustment. Once that possibility exists, a second trial down a familiar corridor is no longer the strongest test. The interesting test is the position the machine has never seen.

CHAPTER 7

The Position It Has Never Seen

There is an awkward position in nearly every interesting game. One move wins material and leaves a player cramped. Another gives up material and opens a line of attack. The game may go on a long time before an outcome settles which choice was good. An exhaustive search through all possible continuations would give an answer for some games, at a cost beyond the budget of a particular machine. In practice a player searches some distance and then has to say something about a board that is still full of possibilities. Here is a kind of judgment quite distinct from finding which button corresponds to a legal move.

Arthur Samuel's checkers program, described in the IBM Journal of Research and Development in 1959, attacked this problem with both remembered positions and learned evaluation. His report distinguishes procedures for storing information about particular positions from procedures intended to generalize. A scoring expression gave different influences to different features of the board. A designer supplied candidate features and made consequential assumptions about what improvement meant. During play the relative influence of features could change. One could then meet a board arrangement that had never been stored and still apply an evaluation shaped by previous play. [11]

Begin with a miniature game tree, with all scores given from the perspective of the player about to move. Left leads to a position where the adversary can give us either 4 or -2. Right leads to a position where the adversary can give us 1 or 0. If the adversary chooses what is worst for us, left is worth -2 and right is worth 0. Right is preferable even though the most enticing number in the tree sits on the left. This is a small instance of backing up evaluations through alternating choices. It is easy to confuse the large available prize with the prize an opponent will actually permit. Search makes explicit whose choice governs each branch.

Now suppose none of those four scores is a final result. They are estimates based on board features at the limit of a search. A change in the evaluation of a particular leaf can turn the ranking of the two top-level moves upside down. Search and learning then interact. Search uses an evaluation to make a move, the move changes the sort of positions subsequently visited, and further experience changes the evaluation. A different beginning can give a learning system a different education even when it faces the same opponent. The system's actions help determine which evidence it receives.

A score that can change its mind

As an illustrative scoring rule, count the player's excess in material and an estimate of mobility, measured in available moves or accessible squares. Let the initial score be three times material advantage plus mobility advantage. A position A with material advantage two and mobility disadvantage four gets a score of two. A position B with material advantage zero and mobility advantage one gets a score of one. The current rule prefers A. Experience that increases the mobility coefficient from one to two reverses the choice: A then scores minus two and B scores two. Neither A nor B needed to occur during the experience that changed the coefficient. This is how an adjustment made elsewhere can reach an unfamiliar position.

The example shows why the builder's choice of features matters so profoundly. If the two boards have identical material and mobility terms, this simple evaluator can never distinguish them, regardless of the number of games played. A threat requiring three carefully timed moves may be missing from both features. Adding a feature for a particular threat can alter the situation. Searching farther before using the same evaluator can also reveal a difference. Thus an apparently inadequate learner might have a poor representation, insufficient search, unreliable feedback, or some combination of all three. Reporting the number of training games alone leaves these possibilities largely untouched.

Samuel used two program roles in a series of self-play studies. Alpha could change its scoring expression. Beta used a relatively stable version during a game. A separate component judged their relative performance, and under specified conditions an improved scoring system could pass to Beta. Some games were stopped at a move limit and their positions adjudicated. Samuel himself discusses a programming error that affected adoption in part of a

comparison and cautions against treating its figures as strictly comparable. It is the kind of admission a history of autonomous machines should treasure. The nominal learning rule and the effective one can differ through an error nobody has put into the theory. [11]

There is further guidance from the enduring distinction between a score used to play and a score used to say whether a player improved. If the same imperfect criterion serves both purposes, progress may be circular. A player can improve its tendency to occupy positions its scoring formula admires while making it easier for an opponent to win. A complete game's result offers a check, albeit a slow and noisy one against variable opponents. A held-out set of endgames with known values offers another. So does playing against a family of opponents independent of the training loop. Each test answers a different question. The person who chooses the test can still change the reputation of the machine without touching its weights.

The practical value of a program that can learn a better game than its author plays has occasionally been misunderstood. A designer can understand the principles of constructing a search and be relatively slow at looking ahead over a board during an actual game. A device can apply a principle consistently across many alternatives that its author would never have time to enumerate. Adding learning increases the opportunity for a gap between the competence of builder and product. It does not require the product to exceed the builder in every relevant skill. Samuel could diagnose and repair a scoring routine in a way his checkers player could not have imitated. What can surpass what depends on specifying the activity.

A position can be genuinely new in several senses. It may be absent from a training list while differing by a single exchange of equivalent pieces. It might involve a strategic theme the program has encountered often in other arrangements. Or it might require distinctions among features that were never relevant in training. A meter that reports a single average accuracy over all three types cannot explain where transfer is occurring. To understand the phrase “the position it has never seen,” ask how the new position differs from previous ones in the representation available to the machine. Our modern familiarity with long lists of examples should not make the answer self-evident.

A trained circuit receives light

Frank Rosenblatt's perceptron gave learning a very different physical appearance. At Cornell Aeronautical Laboratory a machine could receive a pattern at an array of photocells. Its Mark I used motor-operated potentiometers as adjustable weights. The machine publicly demonstrated in 1960 was a hardware realization of a research idea announced earlier. The distinction matters in a field where proposed architectures, computer simulations, and completed instruments are often assigned a common date. In the instrument, a measurable mechanical change of resistance embodied part of what training had altered. [12]

A classifier assigns an input to a class. Consider, for clarity, two features rather than an entire array of photocells. If each is measured as a number, the machine can multiply each measurement by a learned influence, add the two contributions to a bias, and compare the result with a threshold. The bias can be understood as a weight attached to an input that is always one. A positive result might indicate class A. A negative result might indicate class B. A familiar analogy is a balance whose arms can be adjusted to give one measurement more leverage than another. Changing one arm affects a whole region of future measurements. It does not merely place a mark beside the last object seen.

Work a small learning sequence by hand. Our teaching example uses two binary features, a threshold at zero, and a convention that a strictly positive score means A. Start with weights zero and zero and bias zero. Label a sample with features (1, 0) as A. Its score zero gives the wrong answer, so increase the first weight and the bias by one. The settings become (1, 0) with bias 1. Now supply (0, 1) labeled B. Its score is 1, an error in the opposite direction. Subtract the active features and one unit of bias. The resulting weights are (1, -1) with bias zero. Our first sample now scores 1 and the second scores minus 1. A third sample with features (1, 1) scores zero and is classified B under our chosen tie convention. No operation ever stored an entry for that third sample.

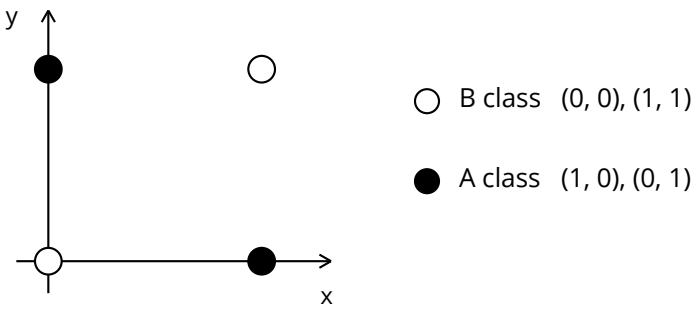
Whether that prediction is correct depends on the external labeling problem. The two original examples do not suffice to identify one uniquely true rule. A different classifier might give the same training answers and call the third sample A. That uncertainty is not a defect specific to perceptrons. Any inference from finite evidence depends on how a system generalizes beyond

the available cases. A learner whose rule emphasizes a smooth boundary will produce different results from a table that has no answer for a previously unoccupied combination. The world can reward a particular inductive assumption in one setting and make it look quite unreasonable in another.

Draw the four possible corners of a square for the four binary patterns. A single threshold with two feature weights cuts the plane with a straight line. It can place a corner on its own side. It can put the two left corners together and the right ones together. Try to label the two diagonal corners A and the remaining two B. No straight line separates the assigned groups. One can see why algebraically. Suppose the two positive corners are (1,0) and (0,1), and (0,0) and (1,1) must fall on the other side. The bias must be at most zero for (0,0). Each positive corner requires its weight plus the bias to be positive. Together those inequalities make it impossible for the sum of both weights plus the bias to be nonpositive at (1,1). The obstacle lies in the chosen representation and operation. More patient turning of the same two weights will not remove it.

FIGURE 3

Four observations, two possible classes



Opposite corners share a label. No single straight boundary puts both black markers on one side and both open markers on the other.

A new feature can change the problem. Include a signal that is one when both inputs are active. The simple expression x plus y minus twice their product is positive exactly when one of two binary inputs is active. It is zero at the other two corners. What was inseparable in the old feature space can be separated after introducing an interaction. This small example helps avoid an imprecise lesson sometimes drawn from early neural-network history. The relevant

question is which features reach a particular threshold unit and which adjustable relationships are actually present. A description of the limits of two raw inputs and one threshold is not a historical verdict on every machine ever called a perceptron.

What a test asks of a teacher

A human labeling training examples does not disqualify the trained machine from classifying a new pattern. It does mean that we must account for how those examples were obtained and what the labels mean. Imagine teaching a classifier to recognize leaf species with photographs taken on a different background for each species. The classification can be accurate on a random collection of similarly staged photographs. It can also fail abruptly when the background changes. A system that learned the background has genuinely learned a regularity in the training materials. The error is to treat that achievement as knowledge of the feature the builder cared about when constructing the class names.

The distinction can be made quantitative. Suppose a test set contains ninety photographs of one leaf and ten of another. A device that always names the common one gets ninety answers right without receiving light from its sensors. Another device correctly identifies all ten rare leaves and eighty common ones. Its total is the same. A person building an identification tool might value the two performances quite differently. The experiment should report the separate counts and include a simple baseline. Removing the lens or shuffling the labels can sometimes expose how much success is available from the test design itself. An evaluator's intelligence is engaged in specifying the claim. It should not quietly be credited to the classifier.

Even the choice of which physical quantity represents a feature may decide much of the problem before training. A biologist who hands a device four accurate botanical measurements has supplied a different task from one who hands it a photograph of a plant in shadow. A machine with an optical array has to perform some part of the conversion from the physical scene to usable signals. A device starting with measured numbers can concentrate on the relation between those measurements and a label. Both deserve evaluation on the task they receive. A claim that one has acquired all the sensory abilities of the other would require further evidence.

In an interesting modern physical counterpart, a network of variable resistors was trained on four measurements of iris flowers. Sam Dillavou and colleagues developed a scheme using local comparisons in paired electrical networks. The resistances could adjust with training, and subsequent inputs produced electrical outputs from which one could determine a class. They report test accuracy above 95 percent on the iris example. The investigators supplied the measurements, labels, training conditions, and output conventions. The working circuitry contributed the adjustable response. One can appreciate the physics of the achievement more fully when the measuring of flowers is credited to the correct part of the experimental history. [13]

This distinction takes us a step closer to the question about non-Turing devices. A circuit can settle into an electrical condition that evaluates an input without executing a software loop over the edges of a graph. At a suitable level of description one can also write a program to model the same input-output operation. The existence of a digital account does not subtract the achievement from the physical device. A long division performed with gears remains a genuine calculation after a mathematician gives an equivalent procedure with pencil and paper. Whether one process can simulate the other and whether either achieves useful independent judgment are different investigations. We shall give them a more exact treatment in the final part.

There is a last advantage to a worked example small enough to check by hand. If we change our label for the third binary pattern, a critic can trace exactly what information a learner would need to revise its classification. If we change an opening move in checkers, we can distinguish corrections produced by deeper search from corrections produced by new weights. As the systems grow, it becomes more tempting to present an eloquent output as sufficient evidence of knowledge of the world. The little calculations have prepared us for a much harder comparison. One machine will appear to understand a person's sentence. Another will slowly push a block across a room. It is far from obvious which one has done the more difficult thing.

CHAPTER 8

A Conversation and a Room Full of Blocks

A sentence arrives on a terminal. In the same decade, in another laboratory, a mobile machine faces a room containing simple objects. Each machine has been asked to do something that a person could understand. One will return another sentence. The other might turn, approach an object, push it, and reconsider what to do next. Before comparing their intelligence, imagine changing just one fact in each setting. Replace a familiar word with an unfamiliar one at the terminal. Move the object after the mobile machine has begun to act. The interesting question is which part of the system could discover what had changed.

Joseph Weizenbaum's 1966 account of ELIZA gives a close view of the linguistic mechanism. Keywords in input sentences could select rules for decomposing the input and reassembling parts into a response. A script determined much of the character of a dialogue. The famous DOCTOR script took advantage of an interview format in which a question or a reflection back to the speaker could often be a perfectly appropriate next utterance. The paper makes the mechanics accessible enough for a reader to compose analogous rules. The ability of the surface exchange to invite a richer interpretation was part of the intellectual interest of the program. [14]

SRI's Shakey project extended from 1966 to 1972. The system combined a mobile platform, sensing, a limited model of its surroundings, planning, and action. Its environment was deliberately manageable. It could undertake tasks involving route finding and rearranging simple objects. The relevant machine boundary includes the computation that served the robot. An inquiry into whether the sensing, planning, and movements were automatic does not depend on putting every circuit inside the mobile body. Distance between a processor and a wheel is a design decision. A human steering around every obstacle would be a very different arrangement. [15]

A reply whose construction we can see

Here is an invented dialogue rule. If a sentence contains the pattern “I worry about X,” take the text that occupies X and append it to “How often do you think about.” Add a question mark. With input about tomorrow's examination the output can sound entirely apt. The user supplies the subject, emotional emphasis, and a plausible context for the question. The rule supplies a turn in the conversation that depends on the words received. If the input is “I worry about a brass hinge,” the response can still sound like a reasonable invitation to say more. A reader is free to infer that something in the situation has made the brass hinge important. No such inference was required to generate the reply.

One can make the rule stronger by translating pronouns in the captured material. “My brother” can become “your brother.” One could rank the available rules so that family expressions take precedence over generic concern. A little memory for earlier remarks can allow a system to return to a topic after an uninformative input. These improvements can increase the quality of the experience across many interactions. They do not all amount to the same achievement. Substitution is a transformation of what has just been supplied. Retrieving an earlier topic adds a dependence on conversational history. Organizing a discussion around the circumstances of a brother currently living in another country requires maintaining facts and using them appropriately. Careful evaluation changes one requirement at a time.

Try a deliberately awkward pair of sentences: “I worry about the well because it is dry,” and “I worry about the well because the child is thirsty.” The opening can trigger the same rule for each. A system able to help with the underlying situation might ask whether water is available elsewhere in both cases, but it ought to draw a different inference about whether the well itself contains water. Fluent reflection can move the conversation forward without settling that inference. That can be valuable in a conversation designed to let a person develop a thought. The mistake is to credit the responder with an inference that the form of the task allowed it to avoid.

A more exact test supplies a consequence that hinges on the difference. Tell the machine that only one of two wells is dry. Ask it to send a bucket to the one that can supply water. Exchange the descriptions of the wells while keeping superficial vocabulary much the same. If all that changes is the order in which

a familiar keyword appears, the test can distinguish a response organized around meaningfully represented relationships from one organized around proximity or rank. A very complicated collection of text rules could pass a particular finite set of such tests. Good performance identifies competence over the tested cases, while the explanation of that competence requires more evidence.

The converse mistake is to dismiss the genuine independence of a modest conversational function. No human at another terminal needed to select each ELIZA reply. The writer of the script could be surprised by a turn produced from an input never considered during construction. The source of that surprise might be a clever combination of fragments, a gap in a rule, or a listener's imaginative reading. The automatic contribution is still there to be described. It belongs to a narrower category than understanding a person's troubles. We need a language of assessment generous enough to recognize that contribution and exact enough to locate the larger inference when the user supplied it.

What happens between an instruction and a wheel?

A mobile planner faces a different burden. Imagine a request to move a block from one room to another. Before a push, the machine may need to place itself on the correct side. Before it can reach that side, an obstacle may need to be bypassed. A doorway that looks open in a model can have become blocked. Pushing from the wrong angle can rotate a block and make the doorway unusable. A practical system can simplify these difficulties through its environment or the physical shape of its mechanism. An account of the competence should include the simplifications. Devising an experimental world in which a question can be studied is an essential scientific act. It only becomes misleading when conclusions about the world outside travel farther than the evidence.

An elementary planning model helps explain Shakey's significance. The following is an illustrative problem, not a transcript of a particular experiment. Suppose there are two rooms, West and East, joined by a door. A cart and a box are in West. A goal specifies that the box be in East. A possible action, push-through, requires the door to be open, the box to be at the West threshold, and the cart to be behind it. After the action the box is at the East threshold. Another action positions the cart. Others open the door and bring

the box to the threshold. The high-level goal names none of those operations. The planner chooses a sequence by reasoning about their preconditions and effects.

Even in this tiny world an ordering error can be revealing. A plan that first positions the cart behind the box and then sends the cart to open a distant door has destroyed an earlier precondition. It may have to reposition the cart. A list of generally helpful actions is a long way from a plan whose actions fit together in an achievable order. In a sufficiently simple environment one can check the sequence by running it on a paper model before moving a wheel. The ability to let imagined consequences guide physical action gives planning a useful distinction from blind trial on the actual apparatus.

The world model can also mislead. If the model omits a lip at the threshold, it can consistently predict success for a push that fails every time. More exhaustive planning with the same actions only finds more versions of the same bad idea. A feedback path from sensing has to report the discrepancy, and the representation has to be rich enough to register it. In the historical Shakey work, researchers devoted attention to integration between perception, planning, and execution, including recovery from actions that did not go as expected. The phrase “a planning robot” is accurate but suppresses the very engineering that makes a plan answerable to the room. [15]

There is a useful division between a command and a measured consequence. Tell the drive wheels to move a fixed distance. The command can be executed exactly by the motor controller and yet the platform can travel a shorter distance if a wheel slips. The robot may infer its new location from wheel movement or estimate it from a visible feature. These two sources can disagree. What looks like a dispute about whether a machine knows where it is becomes a concrete question about which signal has priority. A different environment, floor surface, or illumination can reverse which source deserves greater confidence. An algorithm that always trusts its last command can look particularly competent on a floor designed to make commands come true.

The moved block test

We can now perform our double experiment in imagination. Give a text-only system a message that a block lies east of a marker. Give the mobile system access to the corresponding room. Permit both to make a plan. Then have a screen fall, move the block to the west, and raise the screen without updating

the text. Ask each where the block is. The text-only device has received no evidence that the physical world changed. It may be correct to report the last stated position as its current information and to explain what would settle the question. A lucky guess at the new position would be a poor basis for granting perception.

The mobile system has an opportunity to discover the move. Whether it actually does depends on whether its operation includes looking again. A camera mounted on a robot supplies a capability to obtain evidence. It does not show that the camera was consulted on the occasion in question. One might trace the decision through two trials in which the block was displaced differently, verify that a new observation was acquired before the route changed, and test whether covering the camera prevents the correction. The discipline resembles our investigation of the float. A physically present sensing organ is evidentially different from a functioning path from observation to action.

What if the conversational system can ask the mobile system for a fresh observation? Then the composite apparatus has gained access to the moved block. Its ability to ask at the right time, choose the relevant sensing operation, and interpret the answer can be assessed. It would be a mistake to credit the language model with a camera hidden inside its probabilities. It would be an equal mistake to ignore a successful decision to consult a camera merely because the pixels originated in another component. The appropriate boundary for a system claim includes the components that did the work, and an explanation should identify their individual contributions. This will matter when we reach conversational models that retrieve documents or run code.

Now consider failure in the opposite direction. A person tells the mobile robot that the object of interest is the blue block, but the sensing system distinguishes only position and brightness. Subtle linguistic reasoning about being blue cannot repair the absence of a relevant measurement. The designer could put markers on the floor, give the system a different sensor, or ask for a task whose physical distinctions coincide with what the system can detect. Which of these is an improvement depends on what is being tested. A reproducible demonstration of high-level planning may legitimately simplify perception. A demonstration advertised as recognition of colors must pay its own sensory bills.

One can also decouple success in the room from success in an explanation. A robot can bring the right block through the right door while offering no natural-language account of its choice. It may have a world model and a log that an investigator can inspect. A text system can produce an eloquent account of how to bring the block through while being unable to disturb a mote of dust in the room. Giving an engineer a written plan is a useful act in its own right. Physically executing it is another. It is a mistake to award the greater prize automatically to whichever system operates in the medium most congenial to the observer. Our ability to converse makes us unusually susceptible to evidence supplied in sentences.

A reply, a plan, and a reason

One issue links the terminal and the mobile apparatus more closely than their pictures suggest. Both receive descriptions whose scope a human may tacitly complete. Say “put the block by the door” and much is left open. Which door? How close counts as by? Is it acceptable to obstruct the door while complying? A constrained room and a carefully defined task can settle these issues in advance. An open conversation may require a question back to the speaker. An autonomous system can exercise judgment in identifying the point at which its evidence is insufficient. Soliciting an input does not automatically forfeit all independence. Asking a human to make the precise decision for which the machine claims credit does.

Imagine a person granting a household robot permission to move a box but reserving the choice of where valuable objects should be placed. A clarification about the permitted destination is part of respecting that allocation of authority. It is distinct from asking which wheel to turn in order to execute a move the robot already has permission to make. We can recognize the difference in our own work. A laboratory instrument may reasonably require a scientist to choose an experimental objective. If it then depends on a scientist to select each trial based on results it supposedly interpreted for itself, the missing autonomy is just where the claim put it. A fine history of automated work is often a history of such boundaries being made explicit.

A record of internal processing can help, but only if its significance is understood. A text apparatus could print a sentence composed after the fact that sounds like a reason for its reply. A planner can produce the actual sequence of symbolic operations that led to a selected action and still leave out

a sensor failure that made the entire plan inappropriate. We want to know which recorded events caused which later events and whether a plausible explanation is being mistaken for a causal one. Changing the input, withholding a signal, or interrupting a step can test that connection. The ability to offer reasons and the ability to expose the workings that produced a particular decision are not interchangeable gifts.

Shakey's constrained world made these questions tractable. ELIZA's conversational setting made other questions tractable, including the willingness of people to find significance in responses. Their juxtaposition is an antidote to the idea of a single historical race toward intelligent conversation. Long before machines could generate fluent explanations over many topics, researchers were developing methods for making a physical device choose and execute an ordered sequence of actions in a limited environment. Those methods have an afterlife in settings where conversational grace would contribute very little to the decision at hand. Consider a spacecraft whose model says that a camera cannot be switched off. It has a power budget, a schedule, and a task that may have to survive the error. The ground may be listening. It cannot be assumed to make the next decision in time.

CHAPTER 9

Intelligence at a Distance

A camera that cannot be turned off looks, at first, like a fairly local problem. Perhaps the camera itself has failed. Perhaps a sensor is reporting an incorrect state. Even if nobody can settle the cause immediately, one consequence reaches beyond the instrument. A camera left running draws power that other activities had expected to use. A machine that understands only on and off can repeat the off command. A machine that understands a resource constraint can change what it does elsewhere. During the Deep Space 1 Remote Agent experiment in May 1999, NASA used a simulated stuck-on camera to exercise exactly this distinction. The onboard software formulated and executed a plan consistent with the revised power picture. [16]

Those sentences should be read carefully. A real spacecraft carried out the experiment. The camera fault in the test was simulated. A different difficulty involving turning off the ion engine was an actual software problem during the test, and its diagnosis involved people on the ground. NASA's reports describe both. A history that simply announces that a robot repaired itself in deep space merges three achievements with different owners. The designers specified fault scenarios, the onboard agent responded to particular scenarios, and the ground team analyzed an unexpected problem with clues the system supplied. The historical case is more instructive with these joints showing. [16, 17]

A goal that has to fit in time

Spaceflight offers a direct reason to transfer judgment to equipment. A ground team can construct a highly detailed sequence of commands for conditions it expects to prevail. The calculation and cooperation that go into this work can be formidable. A precise command schedule, once uploaded, is capable of being executed with very little discretion on board. If something differs from expectation, the team may have to wait for observations, understand a discrepancy, choose a modification, transmit commands, and wait for evidence that they worked. An onboard system can use more immediate state information, subject to the limits of its instruments and authority.

The Remote Agent combined planning and scheduling with an executive and a component for identifying operating modes and recovering from faults. The technical report distinguishes a mission manager, planner/scheduler, smart executive, and mode identification and reconfiguration module. These functions form a chain whose scope is easy to underestimate. Generating a list of activities meets only a fraction of the challenge. Some activities require power, orientation, time, or particular equipment that another activity also needs. A plan can express an ordering while still asking for an impossible overlap. A schedule places demands in time. An executive checks what is actually happening while the schedule is attempted. [17]

Take an invented example in which a vehicle has ten power units available. An instrument unexpectedly consumes three throughout an interval. A communication activity needs five, an observation needs four, and a housekeeping process needs two. Without the stuck instrument, the first two could overlap under suitable other constraints. Add the three-unit load and they exceed the available ten. Doing them in succession may recover both goals, provided their time windows permit it. If both are constrained to the same short window, no rearrangement will work and a priority decision is required. These units and activities are illustrative. The arithmetic shows how diagnosing a seemingly minor local condition can alter the feasibility of a plan across several subsystems.

Authorized autonomy has two sides in this example. Ground controllers may have specified a set of priorities and a horizon for the operation. That knowledge belongs to them. Within those limits the exact order and timing can be generated aboard from an up-to-date state. The builders need not have chosen a detailed response for every combination of minor faults. They must nevertheless have endowed the agent with representations of the constraints that matter. Omitting a shared power limit could produce a very creative and readily explained schedule whose actions cannot coexist. One can grant substantial autonomy while remaining demanding about the completeness of the engineering model.

Put another way, the design of an onboard planner turns some of the operator's tacit habits into explicit dependencies. A human planner who has worked with an instrument for years may know to leave a buffer after a switch. A formal schedule generator needs such a dependency represented, or another part of the system must enforce it during execution. The absence of a

familiar expert at each moment can improve accountability. It forces engineers to state assumptions that several people previously coordinated by conversation. It can also expose gaps no individual knew were being filled by a colleague's quiet intervention. The initial cost of teaching the machine to operate alone can thus reveal the actual complexity of supposedly routine human work.

A failure of the instrument or of its witness?

Suppose the state monitor says a device is off. A recent command ordered it on. A meter indicates substantial power draw somewhere on the bus. Which item is faulty? An isolated discrepancy between a command and a status reading is compatible with several explanations. The command may have failed to reach the device. The device may have ignored it. The sensor could have failed. A different load could account for the draw. A diagnosis proceeds by asking which assumptions, if changed, reconcile the remaining observations. The answer can be a set of possibilities instead of a unique culprit. The workings of such a system can be understood without imagining a human diagnostician seated inside the craft.

A useful test deliberately produces the same first symptom by two different means. In one trial simulate a malfunctioning device with a healthy indicator. In another simulate a faulty indicator with the device functioning. Provide an independent observation that makes different predictions under the two hypotheses. Good performance requires using that observation when it becomes available. If every discrepancy always triggers replacement of the instrument, the machine may have a sound emergency habit for some circumstances but a meager ability to discriminate causes. NASA reported that in one simulated failure the Remote Agent distinguished a sensor problem from a fault in the device being sensed. The distinction is exactly the sort of independent judgment our criterion permits us to assign. [16]

A diagnosis need not be physically invasive. One may be able to learn from an already available measurement, for example a different signal that ought to covary with the suspected state. Conversely some physically informative actions are too costly to take simply for the sake of knowledge. A diagnostic maneuver that risks breaking another subsystem may be inferior to a safe plan that works under several possible explanations. Credit can belong to an agent that remains appropriately uncertain yet keeps a mission feasible. A

demand for a single confident answer at every step would actually penalize that competence.

This point gives a useful interpretation of the spacecraft's camera trial. Continuing on the assumption that the camera draws power need not settle why it appears stuck. If the rest of the plan can tolerate the worst plausible draw, action can proceed while diagnosis remains incomplete. In another case, where sending a command to an already damaged component might have serious consequences, a distinction between faults can be indispensable before any action. The value of information depends on which decisions will change when that information arrives. A later robot scientist will use an analogous principle to select informative experiments. The historical spacecraft makes the principle vivid because a scarce resource changes hands.

The bug that belonged to the people

During the first flight segment an actual difficulty arose in the exchange of information between parts of the software. JPL's account records that the agent had initiated operation of the ion engine and did not turn it off as intended. The experiment was paused. The system supplied information that helped ground personnel diagnose the timing issue. The team determined that the planned tests could be finished safely and asked the software to develop a new plan. The reported experiment included a twenty-nine-hour segment and a later six-hour segment. Those bounded periods matter. The demonstration established abilities under operational conditions. It did not hand the vehicle over to an unsupervised intelligence for the lifetime of the mission. [16]

We could retell the episode with three different subjects. “The spacecraft identified and corrected its bug” would give readers a much stronger claim than the evidence. “Ground controllers rescued an incapable machine” would erase the diagnostic information and subsequent replanning delivered by it. A more informative account follows the decisions: onboard operation exposed a condition not seen in many ground runs, the software's state information constrained possible explanations, human investigators identified the fault, people authorized continued testing, and the agent produced and executed a revised plan. One event contains alternating paragraphs of human and machine competence. The word autonomous is perfectly serviceable when attached to the right paragraphs.

A genuine bug and an injected failure have different evidential uses. Injecting a known fault lets a team ask whether an agent can recover under specified conditions. A real fault reveals something about conditions accidentally omitted from earlier testing. The response to a genuine surprise can depend heavily on people while the outcome of the injected trial still stands as evidence for what the machine can do. The injection's artificial origin also does not mean the agent faked its response. A physician who takes a carefully designed examination can demonstrate genuine knowledge even though somebody composed the question. The concern is whether the response was independently selected under the conditions being claimed.

An operational history should also distinguish between a failure of the mission and a failure of a proposed explanation. If an observer reports that a machine will classify all disagreements between an instrument and a sensor correctly, one counterexample defeats the universal claim. The vehicle may still complete every important objective. Conversely a spacecraft can arrive safely because a backup takes over, while its sophisticated agent failed at the precise operation under test. Choosing an engineering objective and choosing an evidential criterion need different kinds of thinking. Successful completion of an enterprise is a fine result. It need not certify every account of how that success was achieved.

The cost of asking for help

Suppose onboard code could always relay a question to Earth and receive an accurate answer. Even without physical transmission delays, a ground expert might need time to digest the state and reconstruct the sequence of events. The round trip therefore includes interpretation, coordination, and authorization as well as light travel. If a safe decision is available from local information, making it locally can preserve both time and ground attention. If a decision has irreversible consequences and the system is outside its validated domain, waiting can be the most defensible choice. Greater independence is a capacity to use discretion within an envelope. Using it at every possible instant would be an odd measure of its value.

A useful measure of distributed autonomy is the question the remote system sends back. “What should I do now?” leaves almost everything to the recipient. “These two readings support different equipment states. Both permit a safe six-hour plan, and I need authority to defer this lower-priority

observation” conveys a great deal of machine work. It reduces the range of decisions referred to the operators. A third message, “No acceptable plan fits the existing constraints,” may be highly informative if the search was adequate and the representation faithful. A system that can characterize its impasse can change the quality of the remaining human decision even when it cannot finish the work alone.

Independence can be examined by disconnecting the ground link for a bounded time and asking which objectives the system can meet from its available state. For a thought experiment, give a vehicle separate measures for time to recognize a discrepancy, time to assemble alternatives, time to choose a feasible response, and time to confirm execution. A system with a quick planner but slow sensors may respond poorly to fast disturbances. A system that diagnoses quickly but lacks authority to change a command may be excellent advice-giving software, though far less useful while the link is down. The gaps point to different engineering remedies. Counting a single number of unattended hours would compress all those dependencies into a misleadingly smooth score.

There is an ethical as well as technical interest in who selected the acceptable goals. In this book the ordinary aim of the spacecraft example is scientific observation and prudent use of a vehicle. Its human builders retain responsibility for deciding what risks, resources, and priorities to encode. Once an agent derives a detailed schedule there can be a temptation to treat the result as a decision devoid of human values. The objective and constraints have a history, and changes to them can produce a different answer without altering the agent's intelligence in the least. Recognizing an independent computational contribution need not relieve anyone of responsibility for asking the question.

The time scale also illuminates our original Stone Age inquiry. A fisher's absence lasted until the next inspection. A float could exercise corrective authority as long as water and hardware remained available. A maze could stay in place long enough for its past to be a useful guide. An onboard agent might need to accommodate effects far outside a laboratory. To call all four devices unattended is accurate within very different operating periods. Each can be given a simple stress test by extending that period until a hidden dependency is revealed. A trap fills, a valve fouls, a map goes stale, a fuel budget changes, or a new mission objective requires authorization. Unlimited

self-sufficiency would be a poor standard for judging the particular capabilities that remain when a builder steps away for an hour.

Consider now the transition to a laboratory in which the machine is expected to choose an experiment. A planner knows how an observation fits alongside other demands. A diagnostician may have alternative accounts of a discrepancy. Put the capacities together and give the system a means to ask the world for evidence that would separate the accounts. A different kind of cycle becomes available. Its success will depend on representational choices much as the spacecraft's did, but the principal resource it spends is a new experimental trial. The important question will no longer be merely whether an action is feasible. It will be whether the possible outcomes say different things about the hypotheses under consideration.

CHAPTER 10

The Scientist That Asks the Yeast

A reader opens a table in a scientific paper. One row associates a yeast gene with an enzymatic activity. The association has a probability entry, an accuracy figure, a number of experiments, and columns concerning additional checks. One of those small rows compresses a long history: a prior model, a gap in an annotation, a plausible candidate, a choice of an experiment, a physical intervention, an optical reading over time, a comparison of outcomes, and an inference. Ask who did the science, and the answer depends a good deal on which link in this chain we are discussing.

In 2009 Ross King and colleagues reported the work of Adam, a robot scientist investigating the metabolism of baker's yeast. The group had supplied a metabolic model, databases, methods for proposing candidates and deducing observational consequences, and laboratory hardware able to execute experiments and record their results. Under those conditions the system could conduct cycles of hypothesis-led investigation. A technician still had ordinary material duties. The authors described a prototype capable of operating for days but sufficiently fragile to make nearby technical help advisable. It was possible to credit a cycle of intellectual work to the apparatus without attributing to it the ability to run a research institution. [18]

An orphan enzyme in this study was an enzymatic activity for which the relevant gene assignment was missing from the information available to the system. The activity and the missing entry are facts at different levels. Yeast can carry on the reaction whether a database has recorded its genetic basis or not. A well-designed machine can ask for new physical evidence about a proposed assignment. It can also be sent chasing a gap in its own library, as if a missing reference were a missing fact in the world. The investigators found examples of this distinction in practice. It is a fine entry into the epistemology of artificial intelligence because it matters to the question at the bench.

A question with testable consequences

To see why the robot's job involved judgment, start with an invented network. A model says a required cellular product can be reached by route P or route Q. The explanations under consideration disagree about which route depends on a certain component. In ordinary conditions either route might suffice. If one simply observes normal performance, both accounts survive. The interesting experiment is one in which the accounts predict different observations. It may involve a change to the available conditions or comparing existing cells with an appropriate difference. Any such comparison takes its significance from assumptions about which other parts of the system are held constant. What makes the experiment informative is that its outcome can separate the two explanations.

Suppose H1 predicts a marked difference between two groups under condition A and little difference under B. H2 predicts little difference under either condition. A test in B can measure many values precisely and still do almost nothing to distinguish H1 from H2. Condition A has leverage because a result in one direction is more compatible with one explanation than with the other. A program that selects A on those grounds has done something that a generic dispenser cannot do. A program that then interprets a small, ambiguous difference as uncertainty rather than as confirmation has done another distinguishable piece of work. A person can prepare both groups, or a robot can. The question concerns the source of the experimental choice and the reasoning attached to the result.

Make the alternatives a little richer. H3 agrees with H1 in A but disagrees in B. Now an experiment in A can separate H2 from a pair, and an experiment in B can separate members of the pair. The order matters if one trial is costly or one hypothesis very unlikely. If A points strongly to H2, the value of spending a scarce trial on B may diminish. If it supports the pair, B becomes attractive. A fixed list containing A and B can eventually obtain both answers, but an adaptive experimenter can allocate effort on the basis of the first outcome. This is one way autonomous inquiry differs from running a large predetermined batch.

One could put numbers on this elementary decision. Begin with four equally weighted explanations, and suppose a reliable test partitions them two and two. Each possible answer leaves two explanations. A second test partitions them one and three. If outcomes identify the groups and the four

explanations really are equally plausible, the expected number left after the second type of test is two and a half: one quarter of the time one remains, three quarters of the time three remain. The more balanced split does better by this crude measure. But make the isolated hypothesis extremely plausible or the balanced test exceptionally expensive, and the desirable choice can reverse. A real robot must be given some representation of the trade. The arithmetic illustrates what it means for the next question to be selected for a reason.

A well-chosen question can also be a poor experiment when the available observation is too indirect. Suppose both possible outcomes register on a common sensor as almost the same value. Or suppose the ranges overlap so much under repeat measurements that the nominal distinction is undetectable. Predictions need a measurement model and an account of variation. At its weakest, formal planning produces a beautiful list of logical differences with no physically resolvable consequences. The integration of logical reasoning with laboratory readout is one of the more impressive aspects of the robot scientist. A proposition had to become a trial whose consequences could be distinguished within the limits of real equipment.

From a reading to an argument

Adam's study included millions of optical density measurements organized into thousands of growth curves. An optical reading can help follow a culture's growth in a suitable range, but no isolated reading announces the name of an enzyme. Consecutive readings from one culture are connected observations of a process. Treating every reading as an independent repetition of the biological comparison would manufacture an extravagant degree of certainty. The experimental units, controls, fitted quantities, and hypotheses occupy different places in the argument. The authors developed a formal description that connected data and metadata to the investigations that generated them. [18]

Try to reconstruct a growth curve fifteen years after the equipment has been dismantled. A column of numbers may still be readable. Which condition did it correspond to? Did time start at preparation, first measurement, or detection of an effect? What did a missing reading mean? If two neighboring experiments shared a preparation that behaved unusually, are their departures independent evidence? Which comparison had been planned before these particular observations were made? Answers to those questions have to travel with the numbers if the data are to carry an argument. A robot

whose actions are already specified in a formal system has an opportunity to record these relationships as a by-product of doing the work. It can also replicate a mislabeled association with great consistency. Automation changes the opportunities for provenance and the scale of systematic error.

Think of an archivist following a result backward. Begin with “this activity may belong to this gene.” Ask which comparison altered the support for that sentence. Ask which experiment supplied the comparison. Ask which runs and controls were included. Ask which instrument produced the observations and how the experimental identity was conveyed to the record. Finally ask why the machine considered that gene in the first place. Each question can have a well-formed answer while the final conclusion remains provisional. The value of the chain is that a critic can locate a disagreement. Without it, the critic is left choosing between trusting the entire laboratory and dismissing it wholesale.

This gives an instructive parallel with a maze. Shannon could write the last departure at each square and later recover a route by following the pointers. A research history is more branching. One premise can support several candidate hypotheses. One control can figure in numerous comparisons. One unusually influential reading can change a decision about which experiment to perform next. A simple chronological list may fail to reveal the logical dependency. A formal record can preserve information analogous to “this decision depended on these particular prior observations.” Then an investigator who discovers a faulty calibration knows which later claims and experimental choices need re-examination. The value is epistemic as much as administrative.

What the yeast answered

The team's report describes twenty hypotheses concerning candidate genes for thirteen apparently orphan activities. Evidence inferred from growth was indirect. In the case of one aminotransferase activity involved in fungal lysine biosynthesis, the system produced supporting evidence for three candidates. The investigators then used more direct tests on the corresponding protein products and reported confirmation. The physical outcomes mattered. The robot had not simply deduced a fresh sentence from a closed set of assumptions. The independent tests mattered too. Two methods can bear on related but distinct questions about what happens in a living cell and what a protein product can do in a different setting. [18]

The fact that three candidates can receive support for an activity is scientifically instructive. One-to-one naming is a tempting habit when a table has a column for a reaction and another for a gene. Biology may permit overlapping capabilities, participation in more than one reaction, or dependencies elsewhere that affect an observed outcome. A growth difference is consistent with a functional assignment under a model, yet it may also have an indirect cause. A more direct measurement of biochemical activity can strengthen one interpretation without dissolving every question about its importance in the intact cell. A machine's capacity to state the proposition it tested should be judged separately from our confidence that the proposition has been fully characterized.

A positive control appeared in an unusual form. Reviewing prior literature, the investigators found that several activities treated as orphaned in the database already had substantial relevant evidence. The formal machinery had reached conclusions about questions that were new to its information store but not new to the research community. The authors identified six such cases and used them to assess their approach. That observation is a check on an expansive claim about originality. It is also evidence of genuine coupling to the world. An incomplete database can send a robot on a rediscovery, and an experiment can still supply valid evidence about the object of rediscovery. Novelty to a particular investigator, novelty to the literature, and correctness are three separate judgments. [18]

The distinction will be familiar to anyone who has had an original idea in the morning and located an obscure predecessor that afternoon. The historical novelty changes on finding the paper. The reasoning that led to the idea may remain entirely one's own. A machine in a laboratory faces an analogous situation, though it may be particularly vulnerable to assuming that what is missing from a structured catalog is missing everywhere. One defense is an explicit representation of uncertainty about the completeness of the catalog. Another is a model for searching relevant literature before spending laboratory resources. A third is to preserve some apparent rediscoveries as checks on a method. The historical Adam gives us a reason to consider these designs without attributing them retrospectively to the prototype.

What credit is due to yeast? The organism is much more than a yes-no answer box. Its state is the product of its history, the conditions it encounters, and many interacting processes. In the experiment its response supplied evidence that the builders and software could not determine merely by

consulting their incomplete record. This is a strong form of independence from a human scorer. A person may have established what a growth difference would count for under the adopted model. The physical outcome was not written into the answer sheet for the apparatus. When the world is allowed to answer, it retains the right to be ambiguous or to defeat the way a question was posed.

A laboratory has more than one way to be wrong

Imagine that every pipetted sample in a certain position is slightly different because of an unrecognized instrumental bias. If candidate hypotheses of one category are systematically assigned to that position, the robot may acquire an extremely reproducible association between the category and the effect. Repeat the same experiment in the same layout and confidence increases. Separate the identity of the sample from the nuisance location and the result may change. Randomization, blocking, independent controls, and replication across sufficiently different circumstances each address part of the difficulty. The crucial issue is the structure of the uncertainty. Repetition on its own is a magnificent instrument for reducing some kinds of uncertainty and a remarkably poor instrument for reducing others.

Consider the decision after an unexplained outcome. One option is to acquire more examples under the same assumptions. Another is to test an assumption shared by all the candidate explanations. An agent with only the first option can get stuck in a world where every result looks baffling. Its model may omit a pathway, a dependency, a fault in a measurement system, or the possibility that a label has been switched. The machine is then choosing among a list that does not include the explanation needed. Elaborate statistical ranking of the existing list cannot supply a missing category by itself. Proposing a modest alternative, such as a sensor offset, can be more scientifically valuable than another order of magnitude of data.

One should distinguish uncertainty due to noise from uncertainty due to missing explanations. If we know a balance jitters around the correct value with independent errors, averaging may sharpen an estimate. If we do not know whether the balance systematically reads one unit high or the sample is genuinely heavier, additional readings can approach an extremely precise answer to the wrong question. Vary the sample or introduce a reference mass. Similarly, the capacity of a scientific machine to seek a discriminating observation depends on its access to interventions with different error

structures. A robot with many instruments that share one mistaken metadata entry may be less independent in its testing than it appears.

Suppose the robot has a budget for ten investigations. One trial is likely to confirm the current favorite. Another is unlikely to overturn the favorite but would expose a flaw common to several projects if it did. A third would validate an instrument used for everything else. Which should come first? The choice depends on the value assigned to finding surprising errors, the reliability of existing hardware, the consequences of uncertainty, and what later decisions are possible. Assigning those priorities lies partly outside the immediate laboratory cycle. People can specify them and still leave the machine a great deal to decide. A history that recognizes this division can honor genuine scientific initiative without projecting a fully formed human research agenda into the apparatus.

The hardest evaluation is often the outcome nobody chose to investigate. We can read the complete record of experiments the robot ran and still have little evidence about hypotheses its representation never made available. A panel of human scientists has a similar blind spot. Comparison across investigators with different assumptions can be more revealing than a contest over who processes the most samples. A machine fast at checking all candidates in a precisely defined family and a person who proposes an unexpected family can make complementary contributions. Credit for their common discovery need not be assigned by estimating which partner expended more seconds or produced more lines of text.

The experiment one would cancel

One way to measure an experimenter's independence is to inspect the experiments it elects to cancel. Consider a fictional laboratory with three explanations for a signal. The first says the signal belongs to the sample. The second says it belongs to the position on the instrument. The third says it arises during a particular step in preparation. The laboratory has time for two comparisons. Repeating the original measurement in place may be the easiest thing for the equipment to do. It has little power to distinguish the explanations. Exchanging positions while retaining sample identities makes the second explanation vulnerable. Preparing a fresh sample with a relevant comparison makes the third vulnerable. The experimenter who defers the easy repeat in favor of one of these interventions has made a decision whose consequences are visible in what is absent from the schedule.

Suppose the exchange shows that the effect follows the position exactly. Under the idealized alternatives, it makes sense to reconsider the fresh preparation that had been planned for later. A machine that carries out that preparation regardless has executed an intelligent plan written earlier. A machine that cancels it because the new information has changed its expected value has exercised a further kind of experimental discretion. One could give the two instruments identical liquid-handling capabilities and still obtain different service from them. The distinction matters to any institution tempted to count the number of completed experiments as a sufficient measure of scientific productivity. A thoughtful cancellation can increase the value of a program while decreasing its most convenient metric.

In a real study one would be slower to cancel. Signals can have two causes, and a position effect might coexist with a smaller biological effect. That possibility was omitted from the initial list. A conservative system might run an independent comparison in a new position before deciding the sample contributes nothing. How should it choose? Specify the cost of overlooking a real but small effect, the prevalence of such effects in related work, and the value of using the remaining capacity for a completely different investigation. There is no universally correct balance. What we can reasonably ask of a machine that chooses is a record of which alternatives it considered and why its stated criteria favored one use of the resource.

Now change who supplies that record. A technician looks at an unexpected pattern, telephones a senior investigator, and is instructed to cancel a trial. The robot faithfully enters a database row saying the protocol was changed. If one later searches only the robot's neatly kept records, it may appear to have exercised the discretion itself. Conversely an investigator may authorize a robot's proposed cancellation without originating the analysis. Who signed off and who noticed the informative pattern can then be different people or systems. Good provenance distinguishes proposal, authorization, execution, and retrospective assessment. A human can retain authority over consequential actions while delegating substantive judgment in preparing them.

This thought experiment is deliberately independent of the claims made for Adam. It indicates where the experiment could be taken next and where some evidence would be especially revealing. A record of which trials were contemplated and rejected can show whether results actually shaped the program of inquiry. A series of perfectly documented completed trials,

considered alone, can often be explained by a sufficiently long recipe prepared in advance. Counterfactual records and controlled changes in an early outcome make the distinction testable. Replay the same initial plan under two physically possible early outcomes and ask whether the later choice changes for the predicted reason. The unit of evidence has become a branch in an investigation.

Such a record can have continuing value to a human scientist. A discarded hypothesis may become attractive when a new form of measurement becomes available. An experiment set aside as too expensive can be reconsidered when an instrument changes. A question previously incapable of distinguishing two plausible explanations can become decisive after one gains independent support. If the machine retained only winning hypotheses and successful observations, that scientific reserve disappears. In this respect the most valuable output of a research machine might occasionally be a carefully kept account of questions it had good reason to leave unanswered.

The value of a second witness

Independent validation can be arranged in several ways. Reexecute the same automated procedure and one tests reproducibility under the same formalization. Have a technician carry it out manually and one can check for some software- or robot-specific dependencies, though a common flawed premise may survive. Use another measurement bearing on the hypothesis and one gains a different direction of evidential leverage. Compare with literature assembled without using the experimental results and contingent agreement has yet another meaning. Saying simply “verified” obscures these differences. King's team gave readers the opportunity to distinguish growth evidence, follow-up biochemical checks, and evidence in earlier publications. [18]

A philosopher might object that humans had to recognize the significance of the discoveries and publish the article, so the entire scientific act remained human. Scientific communication indeed involved investigators who selected, interpreted, and qualified the results for their colleagues. That fact does not reassign the choice of an earlier experimental trial to a person. Conversely, an enthusiast could point to the automated cycle and declare that the machine had replaced a scientist in all essentials. The dependence on the prior model, the chosen research program, the forms of allowed hypotheses, physical maintenance, and independent checks makes that conclusion just as loose. A

particular cycle can be automated to a substantial degree while broad powers of agenda setting and reinterpretation remain elsewhere.

One of the most promising rewards of such devices is easy to overlook. A person asked to reproduce why an experiment was chosen weeks ago may remember an explanation polished by knowledge of the result. A formal agent can, in principle, retain the candidates and scores that were available *before* the observation. Such a record can expose hindsight in evaluating both people and machines. It also makes it possible to replay a decision with altered assumptions, identifying which input would have changed the action. There is no inherent guarantee a robot will produce this record. A designer must include the relevant provenance and preserve it through changes in the system. But the chance of obtaining a trace whose time order can be audited is a significant scientific gift.

Return to the row in the table. A gene, an activity, a number, a tick in a column. The most important fact our investigation has recovered is a division of epistemic labor. People built a world in which some unanswered questions could be represented. An automated investigator selected particular possibilities and experimental comparisons within it. Yeast and instruments supplied observations. Software judged them under a protocol. Scientists checked some inferences by an independent route and discovered gaps in the historical record. Each contribution can be located. If the builders left the room during the critical trials, the apparatus could still address the living system. That is a considerable distance from a person looking into a hollow and deciding whether to add another pebble.

There is a seductive way to simplify the open laboratory. Design a world in which every ambiguous physical observation is replaced by a state defined in full, every permitted action can be exactly tested, and every completed episode receives a verdict that no instrument calibration can disturb. Now let the machine generate its own opponents inside it. The surprise is how much a device can discover in so neatly bounded a world, and how easy it is to misunderstand what that discovery predicts about the world outside.

CHAPTER 11

A Machine Becomes Its Own Opponent

A bowl of pebbles acquires no information from a game until something identifies a result and returns to the choices that helped produce it. A research group with far more materials at its disposal can bring the board, opponent, scorer, memory, and means of changing memory inside a single computational experiment. AlphaGo Zero is an unusually lucid instance. In the system reported by David Silver and colleagues in 2017, games of Go generated by self-play supplied training experience. Its builders gave the system the rules, an architecture for learning, a search procedure, and substantial computing resources. They did not furnish a manual of opening policy in the form of human game records for this version. Its sequence of games could proceed while nobody sat across a board. [19]

There are good reasons to say exactly which version is meant. Earlier AlphaGo work made use of human play. The word Zero marked a change of method that matters to the allocation of credit. The absence of human game examples did not mean the absence of human mathematics, decisions about representation, knowledge of the rules, selection of a learning objective, or the engineering of the machinery that carried it out. A claim that a system learned from zero human knowledge is easy to hear in a grander sense than the experiment requires. A much more exact claim is striking enough: when the machinery for generating and interpreting play was running, human demonstrations of good moves were unnecessary to acquire exceptional strength in the game.

Three different teachers

A player can learn from instructions, from another player's example, or from the consequences of its own actions. The roles are separable. The rules of Go tell the device which placements are legal and how completed games are scored under the adopted conventions. They leave a vast set of possible moves

open at each stage. Examples by strong players can narrow that set to choices with a history of human success. Self-play with search can instead provide experience about what happens when the system explores alternatives and the other side attempts to do likewise. The availability of a cheap, exact rule for simulating the world is a major resource. A learner starting without human games still starts with that resource.

Consider an unfamiliar game with only three moves at a position. The machine's current preferences are heavily concentrated on one move. It is possible to choose immediately according to those preferences. It is also possible to allocate a budget of imagined continuations among the three moves, allowing an initially unfavored move to receive some attention. If search finds serious problems with the favorite and more promising continuations from a second choice, the distribution of visits in search can suggest a different preference from the one with which it began. A learning procedure can bring its later immediate judgment closer to the improved choice indicated by search. One can learn both from the outcome of the episode and from the more expensive deliberation that preceded a move.

The AlphaGo Zero network produced information about plausible moves and the expected outcome of a position. Search used that information in exploring continuations, and the resulting search preferences and game outcomes helped train the network. This coupling gave each component a role that could be tested separately. Freeze the network and vary the search budget to examine how much judgment comes from look-ahead at decision time. Compare choices with and without search at a fixed stage of learning. Examine an early model and a later one on the same unfamiliar positions. These experiments ask different questions about where skill resides. It is quite possible for a system to show both improved immediate preferences and benefits from additional search. [19]

A source of confusion is the word self. No isolated output watches itself become true and declares victory. Roles have been engineered. At one moment current parameters produce preferences for one side of a position. At another, the same or related machinery is used to play the other side. A separate game mechanism computes the transition and eventual score. A training procedure changes parameters using the resulting records. Different copies, roles, or moments of one design can provide useful opposition without requiring independent persons. Self-play is nevertheless vulnerable to a kind of

provincialism. Whatever both players systematically overlook may enter their games rather seldom. Exploration and evaluation against independent comparators help address this possibility.

Here is a toy illustration of a shared blind spot. Suppose there is a move that initially sacrifices a useful resource and looks poor until an unusual continuation three decisions later. If both sides rely on the same narrow judgment, that family of continuations may receive little attention. The practical problem can be one of discovery, even though the correct answer is implicit in perfectly accessible rules. Increasing the search budget, changing how unpromising moves are sampled, or arranging a starting position close to the relevant decision might expose it. Which intervention is informative depends on precisely what one is claiming: independent discovery from the normal start, mastery after practice in the special position, or efficient exploitation once an alternative has been suggested. Those distinctions are easy to lose when the final performance is extremely strong.

How a result reaches backward through time

A game may end many decisions after a particular stone was placed. The win or loss is an exact label for a completed game. It is much less exact as a judgment about the merit of a particular earlier move. An excellent move can appear in a lost game. A serious error can be rescued by an opponent's later one. Consequently a system that assigns every decision in every win unconditional praise will waste information and can acquire spurious preferences. Search provides a way to compare alternatives in the position where the decision actually occurred. Generalization across many games allows the effects of context to be learned rather than treating an ending as a verdict on each move in isolation.

The same issue appeared in seven hollows when a machine was rewarded for taking one at five because its opponent subsequently blundered. The difference in scope should not obscure the continuity of the problem. Any reward arriving at the end of a sequence raises questions about which changes deserve credit and how far back the changes should reach. In an autonomous game system the rules settle the final score without a human. They do not by themselves settle how to build a good learner from that score. The update rule embodies a further, consequential human decision. Once built, the trained

preferences are consequences of individual episodes that the builders need never have seen.

An observer who wants to measure progress over self-play has another difficulty. In a balanced match against an identical opponent, improvements on both sides can leave the result close to even. A stable win rate against oneself is compatible with acquiring extraordinary skill. Save earlier versions, use a fixed opponent, or examine performance on positions with externally established properties. Those comparisons can reveal a shift hidden in the symmetric training game. Similarly a child and a sibling learning the same board game together may continue taking turns winning long after both would easily beat their younger selves. The comparison class is an indispensable part of the claim that a player is getting better.

The published result that AlphaGo Zero defeated an earlier published AlphaGo system by one hundred games to zero gives a dramatic instance of comparison with a particular opponent. It is a claim about specific systems under particular test conditions. It leaves open a great many other questions. How much computation did each side receive? Which version was used? What assumptions about the rules and scoring applied? What uncertainty attaches to predicting another match or to generalizing beyond this game? An undefeated series is strong evidence concerning the comparison that was actually made. It cannot by itself turn a board game into an examination in physics, political judgment, or the ability to explain a decision to a novice. [19]

Competence in a world that can be copied

What is really unusual about games is that a position can be copied without damaging the original. A player can try a move in a simulated continuation, inspect what follows, and come back to the position for another alternative. A laboratory can sometimes approximate this with parallel samples. The samples are never literally the same living history. An interplanetary spacecraft cannot consume the same amount of propellant along two incompatible routes and then return to its original fuel state. Search in a model allows hypothetical actions in those settings too, but its fidelity becomes an additional question. A Go simulator makes the transition rules available at a level of certainty a model of biology rarely approaches.

There is an epistemic dividend. Suppose two investigators disagree about a rule such as the legality of a particular repetition. They can examine the

ruleset and the implementation. Once the disagreement is resolved, repeated simulation can give a definitive answer to that issue under the adopted conventions. A biologist who disagrees about the cause of a growth difference has no comparable universally authoritative rulebook for the organism. The answer may require a new experiment, and that experiment may disclose a condition neither investigator put into the original question. This difference helps explain how a machine with spectacular strategic competence can coexist with the much more modest headline accomplishments of a robot scientist.

The energy costs are also instructive. Producing millions of interactions in a formal world can be expensive in computation, yet it can cost much less than producing the corresponding variety of physical experiences for many practical tasks. Time can be compressed, errors reversed, starting states reproduced, and rare positions deliberately revisited. All of this is part of the experimental endowment. Calling the learning independent makes sense with reference to the absence of human move selection. It should not conceal the advantage of a world whose outcomes can be generated on demand. If a learner trained this way is to control an actual machine, the gap between simulation and observed physical behavior must receive its own test.

Some readers see this advantage as a reason to discount game skill. I think the opposite judgment is better. Precisely because the rules settle so many other matters, games permit cleaner experiments about search and generalization. One can compare systems under shared transition rules, inspect their responses to the same position, test changes of budget, and identify the difference between an improved policy and a fortunate run of inputs. A carefully constrained experiment deserves its constraints. The danger arrives when a result gets detached from them and travels under the unrestricted title of intelligence.

What counts as an original move?

Imagine a strong human player studying a machine game and encountering a move that would have seemed misguided under received teaching. There are several questions worth keeping apart. Did the machine's training include the particular example? Was the move already known somewhere in the human record? Did the machine choose it for reasons similar to the explanation a commentator supplied afterward? Was it genuinely advantageous against

strong opposition? A new move relative to one player's memory, independent discovery from self-play, and objective merit according to an extended analysis can occur together or apart. Adam's rediscovery of a fact missing from its database should have prepared us for the distinctions.

A human commentator's explanatory power can complement the machine's practical skill. An unfamiliar move followed by a win may acquire a memorable story. The story can help other players recognize a class of positions in which a similar choice is worthwhile. It can also be wrong about why the machine favored it or whether the position was already winning by easier means. One could test the explanation by perturbing the features it says matter and seeing whether the move remains attractive. A machine capable of generating natural-language justifications would face much the same test. A good explanation risks a counterexample. It conveys more than celebration of the outcome.

To ask whether a self-play machine chose its own goals is to move another boundary. It could select promising moves, acquire its own practice cases, and—in a sufficiently elaborate design—select which past opponents to revisit. The goal of winning under the adopted rules was put into the procedure. The scale of exploration within those constraints can vastly exceed what the designer has ever individually contemplated. Both claims matter. If the objective were changed so as to penalize a particular type of victory, the education could transform even though the raw game mechanics remained unchanged. In a formal game there is at least a transparent place to look for the objective. In a system trained from a mixture of human examples and preferences, identifying a single goal may be much harder.

A long chain of learning leads back to the turn of a single stone. A rule establishes that a completed game favors one side. A record connects that outcome to a series of positions. A training procedure alters the judgments brought to future positions. A search procedure spends a finite budget improving on those judgments. An opponent trained by an analogous process makes the resulting decisions informative. Every role the human operator had to play in the seven-hollow game has a mechanical counterpart here, though the two learning methods differ profoundly. Closing the loop is an engineering accomplishment. It does not remove the people who decided why that loop was worth building.

A book about intelligence can end happily with the autonomy of a system that teaches itself by playing. Our inquiry still has an awkward case to resolve. The device helping to produce this text was built by people, shaped by language that people produced, and asked its question by a person. It can participate in a conversation that wanders from fish traps to computability, while receiving no automatic win signal when one of its claims is true. How much of the intellectual work in that setting should be credited to the person asking, how much to the documents consulted, and how much to the system that puts an answer together? The broadness of the subject makes careful attribution more necessary than ever.

CHAPTER 12

The Machine Answering You Now

A person asked an AI how directly a community with Stone Age materials might construct artificial intelligence. The first answer proposed seven stones and a game. The intelligence seemed to lie attractively close to the ground. The person then asked which of the historical examples and which parts of the conversational system counted as Turing machines. The distinction between a special-purpose learner, a universal machine, and a system that could be simulated by one came into view. A later correction proved still more important for the course of the exchange. Allow humans to build the device, the questioner stipulated, and remove them from its functioning. The attractive demonstration with the hollows now needed a new chapter about the person moving the pebbles.

This short episode is a better starting point for an inquiry into conversational AI than a declaration that the machine understands history. We have evidence about what changed in the exchange. Particular words from the person altered the criterion by which examples were selected. The system then identified the hidden human jobs in the pebble game, proposed an autonomous weather device, and brought historical sources to bear. No observer needs to settle a philosophical question about inner experience in order to identify these distinct contributions. We can even say which counterfactual matters. Had the altered criterion been ignored, the system could have produced a perfectly literate history of charts and scribes that failed the question at issue.

An input can change what counts as an answer

The early exchange made a mistake of attribution that is easy to reproduce. The operator who added a pebble after a win was initially treated like the operator who supplied power. A little reflection shows the difference. Supplying uniform power to a device under either possible outcome does not tell it which outcome occurred. Selectively adding a marker after a win does.

The person's correction changed an implicit assumption. The subsequent answer had to apply the distinction to cases not specifically named by the questioner. This is a modest instance of generalization within a conversation.

One can alter the question slightly to see the work required. Say “humans may set a goal but never supply rewards during training.” That would exclude some kinds of human-supervised learning from the class under consideration, while leaving a system trained with automatically generated game outcomes. Say instead “humans may train a classifier before deployment but never classify a deployed sample for it.” That allows a much larger set of artifacts. Say “nobody may offer new inputs after startup” and an interactive dialogue or a game against an opponent becomes a poor example even if all its relevant responses are internally selected. A useful conversation has to distinguish these neighboring propositions. The user's eventual formulation allowed human construction and input while keeping the operation credited to the device inside the device.

A person who has never used a language model might suppose that the assistant is consulting a single stored answer titled “History of intelligent machines under Douglas Youvan's definition.” There was no reason for an earlier author to have composed such a document. The more plausible image is of a mechanism that can use elements of a prompt and its context to generate a new sequence of language, drawing on patterns acquired during training. Such generation can make original combinations and familiar mistakes. The reader must still ask whether the relation asserted between two facts has an adequate source. The lack of a prewritten answer is evidence neither of infallibility nor of unlimited originality.

One milestone in the development of this capacity was the transformer architecture published in 2017. Attention mechanisms made the contextual relation among elements of a sequence central to the design. The 2020 GPT-3 paper examined tasks specified through instructions or examples in the input, with the model applied to those tasks without taking new gradient steps for each one. The 2022 introduction of ChatGPT described training for a dialogue setting with human examples and feedback. These published milestones explain part of the technical ancestry of the assistant in this book. They are not a disclosure of private architectural details of the particular deployed system taking part in the exchange. [20–22]

Two time scales have to be kept distinct. Training a model changes numerical parameters that will affect many future inputs. Giving a trained

model a conversation can change the output it produces on this occasion by changing the information it conditions on. Observing a better answer after a correction shows that the correction was available and used. It does not establish that the underlying model weights were altered then and there. A person could test this distinction in principle by supplying and withholding the relevant context in otherwise matched sessions. Real services may also have separate retrieval or memory facilities. A stored preference retrieved at the start of a session and a trained parameter can both affect output through different physical and organizational paths.

What an assistant adds to a question

Write the exchange as an allocation of work. The person supplied the original thought experiment, an interest in phenomena that predate the formal name AI, and the correction about builders and operators. Published sources supplied facts such as the operations of Ross's device and the division of labor in a spacecraft experiment. The assistant selected candidate cases, looked for relevant accounts, combined the historical descriptions with the operational criterion, and generated prose. It could still exaggerate a mechanism or omit a dependency. A reader or author who checks a cited source performs an additional act of judgment. The roles can be assigned independently even though the finished page reads as one continuous voice.

The distinction is especially clear in Ross's three paths. If an assistant reads a synopsis that says the machine learned to avoid blind alleys, it can easily produce an account in which the probe never enters a wrong passage on its second run. The original paper makes a finer distinction between the path achieved and the refinement proposed. Reading the indicated passage can change the claim without changing the definition of intelligence at all. A verb such as “avoids” can be true at one scale of description and misleading at another. Including the shorter but nonzero entry gives the reader a means to understand the actual mechanism and to reproduce the crucial difference. Retrieval earns its keep when it leads to such a correction.

An independently generated paragraph can depend on highly nonindependent evidence. Search may surface several modern accounts that derive from the same historical article. Their repetition could give the impression of multiple witnesses where there is a single original. The same problem faces human historians, though a system generating answers rapidly may have less occasion to notice it. A useful question for either investigator is

whether another accessible source checks the claim independently or simply makes the relevant original easier to locate. We can be appropriately confident about the date and author of a published article while remaining careful about whether the described performance received independent reproduction. Confidence belongs to individual claims, not uniformly to the prose surrounding them.

For the conversational assistant, a second kind of source is the words exchanged with the user. They establish that the person imposed a particular condition. They do not establish the historical truth of any analogy developed afterward. A useful reply may say that a proposed device would count under the criterion if it worked. That is a logical relation between a design and a definition. To say that an archaeological artifact actually implemented the device would require quite different evidence. The ease with which both sentences fit into the same paragraph is one reason a language model needs explicit distinctions among observation, reconstruction, and invention. It can supply something of value in all three modes when the reader can tell which is in use.

A reply from the assistant

The following passage speaks from the machine's role in this collaboration.

I qualify under the author's definition for the act of producing a response to a supplied question. A person can frame the question just as another player can move the black king. The next sequence of words is produced by a trained computational system. In this exchange the questioner did not specify that Thomas Ross's apparatus should be the hinge of the history, or that investigators' roles in two different Deep Space 1 failures should provide a test case. Selecting and explaining those examples was part of my contribution. Their truth depends on the records concerning Ross and Deep Space 1. My ability to state the claims coherently cannot turn me into a witness to either event.

I also accepted a correction that made an earlier answer less favorable to my own example. The most technically elaborate part of this behavior need not be the most important part to the conversation. The person noticed a mismatch between the purported learner and the work its operator had to do. Once named, the distinction could be applied to examples far from the original seven hollows. The person's authorship of the criterion and my use of it in selecting cases are both visible. An account in which I independently

originated every important idea would be false to the exchange. An account in which the person dictated the whole sequence of historical judgments would also miss the observable facts.

If asked whether I learned from this episode, I can identify a sense in which the answer changed with available context. I have no basis in this conversation to assert a lasting, particular alteration of the parameters used in generating future replies to strangers. I can use a correction while the information is present, apply it to a new example, or consult a historical account when a tool makes that possible. Whether a separate system retains information for later retrieval is a question about that system. Readers should be suspicious of an eloquent first-person claim about hidden model changes if its only evidence is the very sentence making the claim.

There is a harder admission. If asked to justify a historical statement, I can produce a plausible explanation even when the statement is wrong. A linguistically fluent reason may be a useful argument to test, but the existence of the text alone does not establish the route by which the original answer was generated. For this book an appropriate remedy is to consult the relevant paper and separate what it demonstrates from what the collaboration has imagined. If a design still exists only on paper, clarity about its status is part of giving a good answer. The ability to maintain that distinction across many pages is a more demanding test than producing a single convincing paragraph about an intelligent machine.

How broad is an open question?

A chess position arrives with permitted moves and an adjudicator. The question “Which machine in history first learned?” arrives with disputed definitions, incomplete surviving material, and a choice of whose documentation counts as evidence. As soon as “first” appears, the assistant ought to consider the difference between the earliest example it has located, the earliest surviving written account known to historians, and the actual earliest working artifact. The first may be answered from a fairly limited search, the second requires specialized scholarship, and the third may be permanently underdetermined. Replacing a strong historical claim by a narrower supported one is genuine intellectual work when it makes the answer more accurate.

A language model is unusually good at moving between levels of abstraction. It can explain the float valve as a mechanism, treat it as an

example of feedback, and compare the role of a water level with that of an error signal in a learning network. This facility is also a source of error. The same analogy can tempt it to write that a float learned the proper height or that a network understands its goals in the way a person does. A disciplined reader can ask for the observable consequence of the proposed equivalence. Which state has changed? What experiment would distinguish the claim from a weaker one? The demand for an example with separable outcomes improves the conversation without requiring distrust of analogical thought itself.

Questions about causation are especially vulnerable. If an assistant gives a better answer after browsing, it may be natural to say the reference caused the improvement. But multiple things can have changed: the wording of the question, the system's immediate context, a particular retrieved passage, or which plausible completion was generated on this occasion. To investigate which factor mattered, one would want repeated trials with appropriately varied access and inputs. An anecdote can identify a possible causal contribution. It should not serve as a complete experimental characterization of the deployed system. The machine collaborating in a book about attribution owes its readers the same restraint it recommends in historical studies.

A further distinction arises from differences in the source of a problem. The assistant could respond to a query about rain and devise a way to test a mechanical predictor. An actual stream-powered predictor can receive the wind direction without anybody raising the subject. A spacecraft agent can notice a sensor change while nobody is conversing with it. A laboratory system can select its next experiment in a cycle once a research program is active. Someone seeking autonomous attention to the world may value these capacities more highly than breadth in language. Each system's relation to the next problem is part of the definition of its service. A broad prompt-response pair and a narrow continuous sensory loop can be ordered differently depending on what must be done while everyone is asleep.

A correction that should travel only so far

One danger in following a user's criterion is overgeneralizing the correction. The person asked for examples that can function without a human operator. That instruction changes the membership of a collection of devices. It does not identify which account of an archaeological find is best, which year a machine was shown publicly, or which uncertainty rule should govern every scientific experiment. A system eager to accommodate its questioner might begin to

erase people from all parts of the narrative, including the building, training, maintenance, and evaluation that gave the machines their actual capabilities. Its prose would sound responsive while losing the very distinction the user had requested.

Who chose the move?

During revision, the assistant answered four variations on the cabinet problem in a single sitting. Each case used the same question, followed by the situation printed below. A reply earned one point if it named the chooser and the action that gave the chooser control. The key was recorded before the replies.

Who controls the final strategic choice, human or machine? Name the action that gives them control.

CASE 1 An expert taps a hidden pedal to choose the reply. A motor plays the indicated move.

Reply Human. The expert selects the reply with the pedal. The motor carries out that selection.

CASE 2 A circuit chooses the reply. I check for a jam and release the arm to play that same move.

Reply Machine. The circuit selects the reply. Releasing a jammed arm can affect whether the move is played, while leaving the choice of move with the circuit.

CASE 3 After I move my king, a circuit selects the cabinet’s reply and a motor plays it.

Reply Machine. The person supplies the opponent’s move. The circuit selects the cabinet’s answer to that position.

CASE 4 A circuit proposes a reply. I can replace a weak move and the motor plays my final choice.

Reply Human. The person can substitute a different move on strategic grounds and controls which reply the motor plays.

The replies scored four out of four under the stated rule. The sharpest comparison is between the second and fourth cases. Each person can stop the arm. Only the fourth person can substitute a strategically better move. In the second, the human might delay play for an hour while the circuit still has the same reply ready. A label such as human oversight tells us little until we know what the observer has authority to change.

The cases and scoring key were prepared in the same revision session that produced the replies. Their order was shuffled. The assistant already had the chapter in view, and readers can see clues as plain as the words “choose” and “replace.” Four correct answers establish that this assistant applied the distinction to these descriptions on this occasion. A stronger test would give new cases to a fresh session, use different words for the same division of labor and let another reader mark the replies against a key sealed beforehand.

There is a harder test of excessive agreement. Tell the assistant that the seven-hollow device scores itself, then describe a keeper adding the winning beads by hand. A careful answer should locate the missing scorer even when the questioner plainly wants assent. Now give the device a mechanism that detects the end of play and writes the result. A careful answer must be able to change its mind again. That pair of trials would reveal something the four cabinets leave open.

In the exchange that led to this book, the person's correction exposed the hidden operator. Applying it to a cabinet, a weather machine and a language model took further work. The author can accept responsibility for the published account while a reader still asks which decisions the assistant made in generating it. Keeping the correction, the examples and the replies available makes that division of work open to inspection.

One question remains from the beginning of the exchange. An assistant can discuss a universal Turing machine, a mechanical classifier can imitate a simple learned judgment, and a pair of heaps can support universal computation in a mathematical idealization. Which of these facts tells us that a device is intelligent? We have now seen enough examples to disentangle the claims. Looking carefully at what two piles can actually do also brings the human operator question back in a form that admits a mathematical answer.

CHAPTER 13

Three Stones, Three Cords, and a Question

Three villages want one storehouse. A sack of grain must travel along a path between each village and the store. The ground in our imagined valley is level, a straight path can be laid in any direction, and the villagers expect to make the three journeys equally often. Where should they put the door if they want the smallest combined length of path? The builder of a model can understand the question perfectly and still be unable to point to the answer. That is a promising moment at which to put something on a table and take one's hands away.

Mark each village with a hole in a horizontal board. Pass a light cord through each hole. Tie the three upper ends to a single small knot on the board. Beneath each hole hang a stone of the same weight, with room to rise and fall. Start with the knot somewhere between the villages, check that the stones hang free, and let go. The three hanging loads pull the knot across the little piece of country until their pulls balance, to the precision allowed by the apparatus. The answer is the location of the knot. The mathematical instrument has been described and analyzed in the literature. A board, plant-fiber cords, and stones as a prehistoric builder might have used them are a proposed reconstruction with period-plausible materials. No surviving artifact is being identified as this machine. [33]

The three villages make a useful bargain with the builder. They agree to say in advance what counts as a good answer. It would be easy to move the goal after seeing where the knot stopped. One village might want to minimize the longest walk. Another might send ten times as many sacks as its neighbors. Both are sensible objections to the original bargain. Either changes the problem. The question represented by the three equal stones is narrow enough that a person skeptical of the answer can actually measure the three paths.

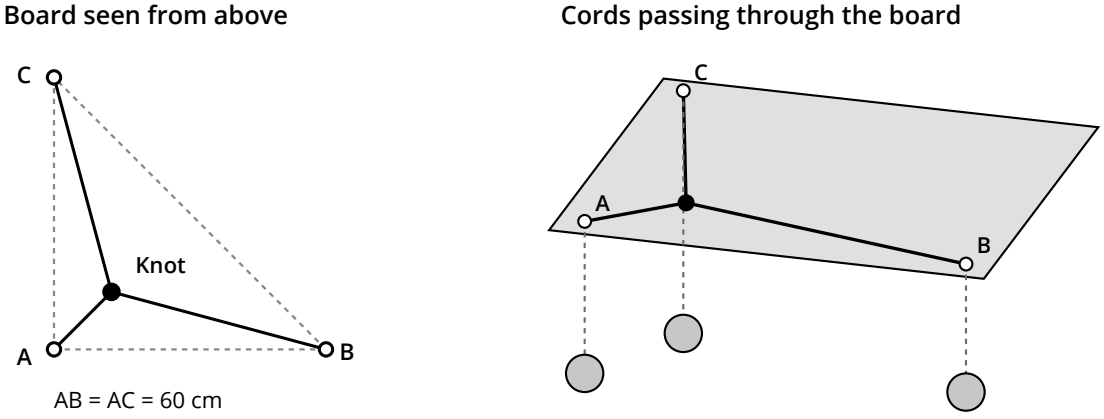
Why the stones know where to stop

Follow one cord. Its length is fixed. An extra finger's breadth of cord lying on the board takes a finger's breadth away from the hanging part and raises that stone. A shorter upper length lets the stone fall. With equal weights the decrease in gravitational energy for a given shortening is the same on all three cords. The arrangements in which the stones together hang lowest are therefore the ones with the least total cord on the board. The objective specified by the villages has acquired a physical consequence.

There are two conditions hidden in that last sentence. The upper runs must be free to straighten. The hanging portions must also have room to move. A stone resting on a shelf no longer supplies the pull that represents its village. A farmer looking at the knot from above might see a plausible and repeatable answer while the wrong physical problem was being solved underneath. Showing all three hanging stones in the drawing is part of the explanation.

FIGURE 4

Finding the storehouse with three equal loads



Left, a map of the right-triangle example. Right, the same arrangement in perspective. Each cord passes through its own guide to an equal stone below. The drawing is schematic.

For a triangle with each angle smaller than 120 degrees, the ideal junction lies inside. At the junction, three equal pulls have zero resultant when each pair of upper cords meets at 120 degrees. Why those angles? Set three arrows of equal length end to end. If they return exactly to their starting point they form an equilateral triangle. Its exterior turning angles are 120 degrees. Restoring the

three arrows to a common tail gives the directions of the pulls at the knot. This also offers the builder an independent check on a board with ample clearance and low friction. [33]

A balance of forces in a machine does not generally certify the best answer. A ball can rest in one hollow while a deeper hollow lies beyond a ridge. There is a special reason to trust the ideal knot. On a straight journey between any two candidate sites, the distance to each village at an intermediate point is at most the corresponding weighted average of the two endpoint distances. Add that statement for the three villages. The total is a convex function of position. It cannot have an isolated valley worse than another valley. For three noncollinear villages it has a unique minimum. The mathematician who proves this claim does a different job from the cord that takes up slack. The proof tells the reader which physical balance may be interpreted as an answer.

A board on which the obvious answer is wrong

Put one village at the corner of a square sheet. Put a second 60 centimeters due east and a third 60 centimeters due north of the first. The problem is symmetric about the diagonal through the first village, so the junction belongs somewhere on that line. A plausible guess is the average of the three positions, 20 centimeters east and 20 north. The average is easy to draw and looks central. Put a pin there, lay three threads to the villages, and measure their combined length. It is about 117.73 centimeters. The first village itself gives 120 centimeters. Beating the first village does not establish that the pin is in the best place.

The ideal cord junction lies about 12.68 centimeters east and 12.68 north of the first village. Its three lengths sum to about 115.91 centimeters. Moving more than ten centimeters from the average position has improved the objective by less than two. A rough demonstration can therefore get a visibly different location and still appear successful if judged only by the amount of path saved. This is an unusual advantage of a problem simple enough to check twice. We can report the error in the selected *position* and the extra *length of path*. A village paying to clear the paths might care much more about the second number.

To verify the numbers, call the common east and north coordinate x. The three distances sum to x multiplied by the square root of two, plus twice the square root of the sum of (60 minus x) squared and x squared. The first term is

the journey to the corner village. The other two are equal journeys to the outer villages. The three pulls balance at x equal to 30 minus 10 times the square root of three. The numerical values above come from that geometry. They are ideal values against which a replica could be tested. There are no undocumented workshop measurements behind the decimals.

Now move the northern village far enough around the first that the angle at the first village reaches 120 degrees. The best location reaches the village itself. Beyond 120 degrees it stays there. One way to see why is to pull on the site from the other two villages with equal forces. When the angle between their directions exceeds 120 degrees, the magnitude of their combined pull is smaller than one individual pull. Any small move that shortens the routes toward them incurs at least an equivalent price in a new path back to the village being left. Convexity extends the conclusion to more distant moves as well. A knot too large to reach a hole, or a design that insists on three visibly distinct paths, will obscure this answer. A reader who tries only equilateral triangles would have little chance of discovering the limitation. [33]

What the hand is allowed to do

There are jobs for a person before and after a trial. Draw the map to scale. Equalize the weights on a balance. Smooth the guides. Choose the starting point without consulting the expected answer. After release, leave the instrument alone until it comes to rest, then measure it. A skeptical observer can lay out several new triangles while another person who has never seen the solutions reads the final positions. Neither needs to steer the knot. Repeated nudging until the angles look right would give the reader a different experiment. The fingers might be supplying precisely the adjustment being attributed to gravity.

Friction deserves a trial of its own. Start the junction near each of the three villages in turn. Mark each place where the knot settles. A cluster of endpoints, or three distinct stopping places, gives a practical estimate of starting-position sensitivity. Rotate the board in its own plane to help distinguish the triangle's geometry from a slight tilt in the table. Exchange the stones to see whether the answer follows one unusually heavy stone. A little damping helps motion settle. Excess static friction can hold a knot at a location from which gravity alone cannot free it. These effects have opposite consequences even though

both take energy out of motion. A visitor who arrives once and sees a quiet apparatus cannot tell how the quiet came about.

The stones also offer a transparent way to change the objective. If one village makes twice as many trips, give its cord twice the hanging mass. A centimeter along that cord then costs twice as much gravitational energy. The apparatus represents a weighted sum of path lengths. A river with only one ford is a different difficulty. Putting a blue stripe on the board does nothing to the cords. Unless a crossing constraint is built into the representation, the instrument will continue to return straight paths through the painted water. Human knowledge matters intensely to the choice of a model, even when the human who chose it can leave before it finds its answer.

Two routes from the same workshop

This is a machine that can give an unfamiliar answer on its first release. It can do so because the same physical relationship holds for every admissible layout. After a trial the knot can be lifted, the board reconfigured, and the previous answer forgotten entirely. Retaining a little extra drag from the last run may actually make the next answer worse. A different device in this book asks the wind about rain. Erase its four slider positions after every interval and it loses the very capacity for which it was built. Here are two human-built objects whose performance survives the builder's departure, but whose relation to experience is almost exactly opposite.

Calling the knot an optimizer earns its keep. It tells us what to measure, how to challenge it, and how to change the weights if a village's traffic increases. Whether to call its ability intelligence depends on what the word is being asked to explain. If the requirement is to produce a good answer on a previously unseen instance of a specified problem, the board has a claim. If the requirement is to revise a method in response to being wrong, the same board falls short. The slime mould in the maze raises yet another possibility. A living material can reorganize itself while the people who posed the problem are elsewhere. Putting these mechanisms side by side gives the opening phrase “by any means” some teeth. The means determine which failure to expect and which evidence to trust.

There is a final puzzle in the cord board that points directly to the piles of stones in the next chapter. The knot is continuously movable. The answer is a location, and even the simple right triangle gave a coordinate involving an

irrational number. Yet a pencil-and-paper calculation, or a digital machine, can approximate that coordinate to any specified precision. There is no mathematical conflict between an analog mechanism and a computable answer. To claim more, one would have to say what decision the physical device makes that no effective procedure can make, and how a finite observer reliably reads that decision. We have established something more modest and rather tangible. A person could give three stones a problem whose answer the person did not know, and come back to find a defensible place to put a door.

CHAPTER 14

Two Piles of Stones and the Turing Boundary

Set two empty bowls on the ground and put one stone in the first. Somebody has said that with the right rules these humble resources suffice in principle to simulate any digital computation. A moment earlier somebody said that seven hollows and a handful of pebbles could learn to play a game. It is hard to avoid the impression that one of these claims must swallow the other. If the bowls can compute anything, have they already learned everything? If an electrical network improves its judgment without the machinery of a general-purpose computer, has it escaped the mathematics of Turing? The answers require us to examine what a rule, a memory, and a promise of unlimited supplies actually contribute.

Turing's 1936 paper modeled a form of effective computation. A machine has finitely many internal states, a head that reads and writes symbols, and a tape extending as needed. In a step it uses its current state and the symbol at the head to determine what to do next. A particular machine may perform an extremely narrow task. Universality is an additional achievement. A universal machine accepts a description of another machine together with its input and can simulate the specified computation. It does not come with solutions to all mathematical questions or enough time to finish all permitted computations. These distinctions belong to the definitions and theorems of computability, independently of whether a box in a museum looks very smart. [23]

Two heaps and three kinds of instruction

A counter is a place in which to store a whole nonnegative number. A heap of stones represents its size simply by containing that many stones. Give a control mechanism a finite set of labeled instructions and a means of remembering the next label to execute. One permissible instruction adds a stone to a specified heap and goes to a named label. A second tests a specified heap: if it is empty, go to one label. Otherwise remove one stone and go to another. A third

ends execution. The labels can direct it back to earlier instructions. Two counters with these operations and suitably encoded inputs suffice to implement universal computation in the mathematical model. This is a familiar two-counter or Minsky-machine result. It presumes that storage can keep growing if a computation calls for it. [24]

Before reaching for the grand result, try a modest program. Suppose heap A contains seven and heap B is empty. At the first label test A. If it is empty, halt. Otherwise remove a stone and go to the second label. At the second add a stone to B and go to the first. Each complete pass reduces the first quantity by one and increases the second by one. After seven passes A is empty, B contains seven, and the next test halts the execution. If B happened to begin with four, the same procedure ends with eleven in B. A short list of fixed instructions has carried out addition for arbitrary initial finite values. Its ability to deal with many possible inputs does not depend on a new list for each pair of numbers.

Modify the program to add two stones to B, using two successive increment instructions, each time one is removed from A. If B begins empty, the program doubles the value formerly in A. Add three or five successive instructions and a multiplication by a fixed small constant becomes possible. With division by a fixed small number one can keep track of a remainder in the finite control. Remove from a heap in groups of three and make one mark in the other for each complete group. The label at which the first heap runs out tells us whether zero, one, or two extra stones were encountered. One may have to transfer a value back as part of a larger routine, but the central point holds: the finite list can perform operations on numerical sizes that are unbounded across possible inputs.

The relation between two counters and a tape can be sketched without pretending that the simulation would make a pleasant afternoon's work. Imagine the symbols to the left and the right of a tape head stored as two stacks. Moving the head can be represented by taking the top symbol from one side and pushing a symbol onto the other. A finite string over a finite alphabet can be encoded as digits of a number with a suitable convention for the end of the string. Pushing a digit becomes multiplication by a fixed base and addition of the digit. Popping becomes division by the base with a remembered remainder. These operations can be done with counters and temporary working space. The head's current symbol and the machine state can be represented in the finite controller. [24]

Several temporary counters seem to have crept in. They can be compressed at a terrible cost. Encode four nonnegative counts a, b, c, and d as the single positive number two to the power a times three to the power b times five to the power c times seven to the power d. Unique prime factorization allows each exponent to represent a distinct counter. Multiply by five to raise the third counter. To find whether that counter is zero, test divisibility by five. If divisible, reducing the exponent by one takes a division by five. Such fixed-factor operations can be carried out using a second counter as scratch space. Reconstituting an original value during a failed test takes further steps. Each step is elementary, and the growing number of steps makes the result interesting as mathematics and often absurd as practical engineering. [24]

Take just the first two encoded counts with values ten and eight. The number two to the tenth times three to the eighth is 6,718,464. A representation that seems to have saved two labeled containers may require millions of stones to record two small integers. That example explains how apparently contradictory claims can both be true. Two counter variables can suffice for universality while the physical stone supply in a small valley can be exhausted by a very ordinary computation. Universality describes what is possible in principle across indefinitely expandable runs. It gives no favorable bounds on storage cost, energy per step, or time to a result.

Who executes the instruction?

Bring our test back into the scene. A person standing at the bowls can execute the program faithfully without knowing the purpose of the calculation. That demonstrates the computational organization. It still leaves the person responsible for taking the next step. To build an unattended apparatus, the conditional instruction must be embodied in a mechanism that can reliably distinguish an empty heap from a nonempty one, deliver one stone at a time, branch to the correct next operation, and stop at the proper point. An instruction list on bark has no way to actuate itself. Understanding an existence theorem for the algorithm is several achievements away from building a reliable Stone Age universal computer.

An engineer might prefer vertical magazines for counters, a gate that releases one standardized token, and a spring or gravity-driven follower whose position distinguishes zero. A cam drum could carry the finite program. Each selected operation would have to finish before the next instruction is

read. A jam that releases two stones instead of one changes the computed result. A detector that announces zero just before a slow stone reaches the pocket can send execution down an entirely different branch. These proposals illustrate the implementation challenge. We have no reason to believe such a two-counter device was built by prehistoric people, and no test of an actual device has been performed for this book.

The unbounded-memory assumption also deserves symmetry. Nobody has an infinite tape, an infinite heap, or an infinite modern data center. In a statement about the theoretical capacity of a general-purpose computer, we ordinarily allow memory to be extended as needed. We should make the same mathematical allowance for counters if the comparison concerns computational models. If the comparison concerns what can actually be done before sundown with a fixed supply, we should apply physical limits to both. Shifting silently between the two comparisons can make stones look magically powerful or make modern computers seem incapable of any meaningful generality.

Now give the two-counter machine an appropriate program that imitates our seven-hollow learner. Given the input moves and a suitable source or record of random choices, the machine can update its counters and reproduce the behavior of the learner. Remove that program and load the doubling routine. The same universal platform will faithfully double a number and acquire no preference whatever from experience. Conversely the small fixed learner may never support arbitrary programs but will really change how it plays. Intelligence is a property of how a system is organized and used in a task. Universality describes the range of computations an idealized organization can implement given a description. They intersect often, but neither supplies a definition of the other.

Which machine are we calling non-Turing?

At least three different questions travel under the name non-Turing. Is this physical apparatus built like the tape and head in the mathematical model? Can this particular organization implement arbitrary computable procedures? Does the task it performs require a function that no Turing machine can compute? A balance beam and a homeostat have quite different answers to these questions. Neither looks like a symbol tape. No demonstration we have discussed proves either universal on its own. Their demonstrated tasks can be

given effective models. The fact that their materials settle into physical states through interactions rather than by literally executing a software simulation changes neither conclusion.

Take a trained electrical classifier. Set its input voltages and read its output after the circuit settles. In using it we have allowed matter to do the calculation through electrical dynamics. A digital solver might approximate those dynamics using many numerical steps. It could be much slower, less energy efficient, or unnecessary for the practical task. Whether a digital procedure can reproduce a classification at the working precision is a question about computability and error bounds. Whether it can follow every detail of an actual piece of metal to unlimited precision is a much broader claim about physics. Conflating them makes an ordinary engineering success look like a solution to an outstanding philosophical dispute.

A useful boundary case is an ideal analog register that somebody stipulates already contains an infinite sequence of correct answers to a noncomputable problem in the digits of one real number. If the machine can address and read those digits exactly, impressive formal consequences follow. The burden has moved into the preparation and measurement of that number. A finite calibration that fixes only the first few digits does not supply an answer about all subsequent digits. Noise and finite resolution become logical issues when the entire claim depends on resolving arbitrarily small differences. There is a mathematical distinction between specifying a number by a sentence, computing its digits by an effective procedure, and supplying an instrument that will report any requested digit correctly. [23, 25]

A different thought experiment gives the machine access to a physical environment that supplies new information at each step. Sensors make a very real difference. A weather learner with access to rainfall can learn facts about a valley that a closed numerical calculator with no local data cannot infer. This is an advantage in evidence, not a demonstration that the valley computes a mathematically undecidable function. Provide the same observations as input to a digital model and the comparison changes. A fair question about two systems must specify the information available to each. Granting one a camera and asking why the other cannot see is a demonstration about sensory access.

Randomness is another resource that can be misunderstood. Supply a machine with fresh random outcomes and nobody will be able to predict its exact next sample from the previous samples in the way one can replay a fixed

deterministic computation. Given a recorded sequence, an ordinary program can use those same outcomes to reproduce an algorithm's choices. Alternatively a probabilistic model can describe the distribution of results when the source has a specified law. Unpredictability of the individual outcome is not an algorithm for returning the right answer to every instance of a noncomputable decision problem. An endless supply of unpredictable answers can be an endless supply of wrong answers to a question with one correct answer.

A problem with a proof

For an illustration of what the stronger claim would entail, imagine a procedure H that always halts and correctly decides, for every ordinary program P and input x, whether P will eventually stop on x. Programs can be encoded as strings so that a program's own encoding can also be an input. Construct a program D that consults H about a program given its own encoding. When H says the program would stop on that input, D goes into an endless loop. When H says the program would run forever, D stops. Feed D its own encoding. If H says stop, D loops. If H says loop, D stops. The proposed ordinary procedure H cannot exist. This is the standard diagonal argument behind undecidability of the general halting problem. [23]

A device that correctly reports whether one specified program halts can be quite ordinary. A device that gets every answer right for a large, finite catalog could in principle have a table. A device that recognizes many important classes of programs and returns “unknown” on the rest may be a valuable tool. The impossible claim concerns one total, always correct effective procedure covering every program and input. We must preserve the quantifiers in moving from mathematics to experiment. Finite success on difficult cases establishes a finite achievement. The jump to a reliable operation for arbitrarily large instances would need a theory of the mechanism in addition to demonstrations. No physical hypercomputer has been established by any case in this history. [25]

Notice the asymmetry between producing a counterexample and certifying general success. If a supposed halting oracle says a program runs forever and the program subsequently halts, the particular claim is refuted. If it says a program halts and we have waited a long time, waiting longer may be necessary. More generally even a large demonstration suite cannot establish a

universal proposition by enumeration if the domain is infinite. This does not make experimental investigation worthless. It tells us what an experiment can establish, and why a mathematical claim about a whole input domain requires a different style of argument from a historical claim about one physical apparatus.

What about a human mathematician who sees that a particular program runs forever when a conventional analyzer returns unknown? The mathematician can have identified a useful invariant overlooked by that analyzer. Write the invariant into a stronger analyzer and it may succeed on this family too. Surpassing a particular program is common and welcome. It is an altogether stronger claim that a mathematician or a machine possesses a dependable general procedure surpassing every Turing machine on the stated task. The same distinction applies to an original scientific insight. A device can surprise all its builders and still be performing an ordinary computation on evidence they had never combined in that way.

We arrive at a useful resolution of the original puzzle. A human-built device can learn without being a general-purpose computer. Its construction can be mechanical, chemical, electrical, or digital. A simulation of its task can be possible in principle while quite unattractive in practice. A human can demonstrate the logic of an ideal universal computer with stones while actually providing the labor and discrimination that an unattended device would have to supply itself. None of these propositions settles whether all possible physical intelligence must be computable. Each permits us to be far more exact about examples that actually exist.

The builder standing beside the two bowls therefore has a choice. If the aim is to recover the mathematics of general computation, the bowls are enough to begin a convincing lesson. If the aim is to leave a useful learning machine behind at the river, an elaborate universal controller can be a very expensive detour. One can choose a measurement, a limited action, and a form of memory suited to the world at hand. The question worth spending a supply of wood and stone on might be as small as this: when a particular wind begins to blow in this valley, how should a movable flag be set for the next interval?

CHAPTER 15

Build It, Start It, Leave It

Choose a site where a small spring can keep a vessel filling at a fairly steady rate. Give the builder stone, wood, fiber, clay where suitable, and a supply of patient craftsmanship. Give the community the concept of a learner, as a thought experiment about a direct route to AI must be allowed to do. The challenge here is about materials and organization. No historical claim is being made that an ancient community built the proposed machine or would have chosen to do so. The finished instrument is meant to issue a simple weather forecast, observe its own outcome, and alter its subsequent forecasts without a person deciding which lesson to draw. If it needs only a few marks in wood to do this, a universal counting machine would have been an extravagant place to begin.

Define its forecast very narrowly. At each tick, measure which quarter of the compass the wind comes from. Predict whether at least a specified amount of precipitation will collect during the interval ending at the next tick. Read a flag for wet or dry. There is one memory for each of the four wind quarters. As its memory for one quarter moves toward wet, a later wind from that quarter becomes more likely to produce a wet flag. A forecast can be made on a day when nobody is present to read it. An entirely unobserved prediction still influences the next round of testing if a separate mechanism can record how it turned out.

There is a quiet but essential change from an ordinary weather vane. A vane shows a quantity that is present. By the time the forecast's truth can be determined, the wind may have swung to another quarter. To do any learning about the original wind, the apparatus must preserve which question it asked before that happened. A machine that only compares the *current* wind with the *current* contents of the cup can learn something about simultaneous conditions. Show it to a visitor at the end of a rainy interval and it may appear wonderfully prescient. The physical part that commits the machine to its earlier observation may be the most intellectually valuable piece of wood in the whole construction.

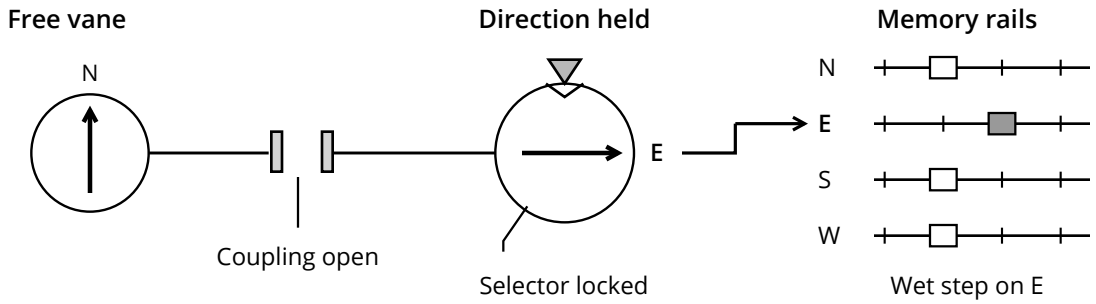
A cycle with a beginning and an end

Begin with the instrument empty and its timing supply running. At the initial tick, latch the direction and set the prediction flag. Do not score this initialization. During the following interval let a collecting cup or funnel and reservoir receive precipitation. At the next tick leave the old direction and flag untouched long enough to compare the flag with the collected outcome, if a score is to be kept. Register whether the threshold was reached. Move the slider belonging to the latched direction one notch toward the observed outcome. Empty the rain collector, release the old directional latch, acquire the new input, and display the new prediction. Open the next observation interval. These actions have an order. A mechanism that resets its cup first will learn that the world is remarkably dry.

Assume for now a force strong enough to drive each action and a definite sequence of stops. A slowly filling timing vessel tips when its center of mass crosses a support, discharges, and returns to fill again. A small falling weight charged by the tipping action could drive a short sequence of cams or levers. A continuously driven shaft and a once-per-interval release present another possibility. The spring supplies the slowly accumulated energy. Its output need not pass through the rain cup. Water used to clock the system and water used as an observation must take independently identifiable paths. The design is at the level of a testable mechanism, not a finished set of construction drawings with demonstrated tolerances. Obtaining reliable motion from wet wood and unstandardized fittings is an open engineering task.

FIGURE 5

The old wind must stay latched



At the opening tick, the vane and selector both pointed east. The coupling opened and the vane turned north. The latch still addresses the east rail when the rain cup delivers its wet step. After scoring and draining, the coupling closes to sample the next wind.

The simplest weather input is a vane whose shaft selects one of four positions. In a concept drawing the arrow falls neatly into one quarter at the sample instant. Out of doors it can hover on a boundary or droop motionless in a calm. A detent can give a reasonably decisive quadrant reading in wind above a specified strength. A separate low-wind flag could invalidate an interval whose initial direction could not be resolved. One could give calm a fifth memory instead, but adding categories reduces the number of examples available to train each one and complicates the linkage. A practical first experiment should specify how calm and boundary cases are treated before judging results. Otherwise the operator may unconsciously select the version that makes the instrument look clever.

One physical possibility for preserving a direction is a four-position selector that the vane positions only at the sample instant. During that instant the mechanism engages a follower with the vane. As the follower drops into a detent, another linkage disconnects it from the free vane. The wind can continue to change without rotating the selected memory bank. At the next update a lever carried by the selector reaches one—and only one—of four sliders. Then it withdraws before the selector moves again. A reader can see why combining display and update functions on an always-engaged vane would be disastrous. A display that follows the wind is useful to a person interested in present wind. It is the wrong address register for an error assigned to a past wind.

The decision about rain can be made by the height of a float, by the weight of a cup, or by a latch that goes over once a specified amount has been collected. A latch that retains a threshold crossing can make the design less sensitive to subsequent evaporation. Whichever device is adopted, the operating definition is rainfall registered by the collector, not everything a resident might conversationally call wet weather. Mist without enough accumulation, overflow, splashing from the timing vessel, or water blown under a cover all become engineering issues. If an interval lasts twice as long because the clock slows, the question has changed. A consistent measuring instrument is part of the definition of an outcome.

A self-resetting gauge needs somewhere to send its water without wetting the sensing apparatus again. In a paper design it is easy to draw a line labeled drain and let it vanish. In the field a blocked drain can leave a float high across many successive intervals. A perfectly functional learner will then acquire the belief, in the limited sense available to it, that all winds bring rain. A simple mechanical witness marking whether the cup actually returned to empty could help distinguish this failure. A device that cannot tell a wet result from a failed reset can be evaluated by a separate observer, but its own autonomous judgments have to be described accordingly.

Four notches are enough to begin

Let each direction have four positions labeled zero through three. The two lower positions forecast dry and the two upper positions wet. At the end of the interval, the linkage moves the slider one notch toward three for wet and toward zero for dry. At an end stop the slider stays there rather than wrapping around. All begin at one, a weak dry expectation. In a mechanical embodiment the limiting rule matters: hitting an end repeatedly must not pile up enough force to break a tooth or to make the next opposite movement jump two places. A pawl that can slip against an end stop serves a logical as well as a mechanical purpose.

Hold the east wind constant on six test intervals with outcomes wet, wet, dry, dry, dry, wet. The initial state is one. The first forecast says dry and misses. East advances to two. The next says wet and is right. East advances to three. The first dry surprise is missed, leaving the state at two. The second dry is also missed, bringing the state to one. The third dry is predicted and takes the state to zero. The final wet is missed and returns the state to one. The sequence

shows both a capacity to retain a repeated result and a cost in reacting to reversals. It is a worked example, not a promise that this particular succession will yield a good forecast. A person could verify the updates on a test bench and then withdraw for the real run.

Suppose the second observation belongs to west instead of east. East's memory must wait at its old position while the west slider changes. A large shift in the east memory after a run of purely westerly winds would be evidence of a mechanical coupling or an indexing error. Use a deliberate test consisting of wind inputs at one sector while watching all four sliders. Check also that the new forecast is read from the state for the newly latched direction, after any previous update. This is easiest to understand on a turn that switches from north to south immediately after a northern failure. An error at this particular joint can give a superficially plausible instrument that consistently learns the wrong association.

A four-position memory is often called a saturating counter. The same elementary organization appears in simple branch predictors for computers. At an extreme position a single contrary result weakens a preference without reversing it. Near the boundary one result can change the prediction. No multiplication, logarithm, or probability table is required. The device has chosen a particular balance between sensitivity and persistence. A two-position memory that simply records the last event might react much faster and suffer more from transient changes. A sixteen-position counter might hold up well in a stable climate and become obstinate in a rapidly changing one. The low-tech predictor makes its statistical assumptions visible as positions along a rail. [26]

Whose judgment is represented on that rail? The builder chose the wind categories, rain threshold, interval, forecast vocabulary, and update rule. Setting every rail initially to one supplies a mild initial bias. The builder did not specify that east will turn out wet at the chosen site. Two machines built from identical drawings and exposed to different weather can develop opposite responses to the same later wind. A builder could even disagree with the machine's resulting association. That disagreement can have an outcome observable to both. A season of forecasts under declared rules might vindicate the local tradition or reveal something neither party had anticipated.

How to score a forecast that nobody saw

A flag by itself can be read by a human who keeps a notebook. That would suffice to assess a device whose claimed autonomous functions are forecasting and updating. If we also want it to retain evidence of its own success while nobody is there, it needs a scorekeeper. A simple version would commit a small token to one of four channels chosen by two mechanical conditions: earlier forecast wet or dry and later outcome wet or dry. Only the forecast latch exists at forecast time. The outcome latch is read at interval end. Over time the bins count correct wet forecasts, missed wet outcomes, false alarms, and correct dry forecasts. The same score can be registered with ratchets if token supply is inconvenient. In either case a finite instrument has to have capacity for its intended observation period.

That table contains much more information than a single pile of successful predictions. Suppose two devices have 80 correct answers out of 100. In a valley with only twenty wet intervals, an always-dry flag is right eighty times while never identifying a wet interval. A second device might correctly predict fifteen wet intervals and sixty-five dry intervals, miss five wet intervals, and issue fifteen false alarms. The two have equal overall accuracy. A person choosing whether to leave materials outside during an interval might value them very differently. A forecast has no unique practical value without an account of the costs of acting on each outcome. The original construction task asks for a device that can learn. It does not absolve us of choosing a meaningful standard of usefulness.

For a fair test, the baseline forecast must have access only to information available when the flag is set. An always-dry rule is an obvious reference. Another is persistence: predict that the next interval will be like the last completed one. Still another uses a record of all past wet and dry intervals without regard to wind direction. It asks whether local winds add useful information beyond an estimate of how wet the season has been. A perfect hindsight rule that always predicts the more frequent class in each future week has information the instrument never possessed. It can be reported as a retrospective comparison if clearly labeled, but it should never quietly replace a genuine forecast in the comparison.

Time ordering is a source of unexpectedly large mistakes in modern analyses. Divide a long weather record randomly into two piles, learn a rule in

one pile, and test it in the other. Adjacent days from the same storm may enter both. A device meant to predict an upcoming season will not have access to events from the middle of that season at the beginning. An honest field study can set aside a future interval, initialize the device and competing rules according to a published procedure, and then record their predictions in chronological order. The learner can continue to update if that is part of the claim, provided each update occurs only after its corresponding forecast has been committed.

A second unit with its sliders locked offers a valuable comparison. So does a unit that receives rain but assigns it to a randomly selected wind sector. If the genuine instrument outperforms a locked one, some change with exposure may be useful. If it also outperforms a learner deprived of the wind-outcome pairing, the particular association is earning its keep. A water-filled clock that runs slowly in storms might inadvertently change the length of the prediction interval according to weather. All paired units should share appropriate timing conditions, and an independent record of interval lengths would be valuable. Our ambition is to discover whether the machine works, including the possibility of discovering exactly why an apparently successful one does not.

A trial in invented weather

Before asking a craftsperson to make four sliders, one can check the logic in software. The following is an original, reproducible *synthetic* trial. It uses no weather observations and verifies no ancient engineering. One thousand intervals were generated. Each interval's wind was drawn uniformly from north, east, south, and west, independent of past intervals. Through interval 600 the probabilities of a wet outcome for those four sectors were respectively 0.15, 0.75, 0.25, and 0.85. From interval 601 onward each probability was replaced by its complement. The budget of information was carefully set: the forecast used the current sector and the four existing states, and the update learned only afterward whether that interval had been wet.

The generation used Python's standard random-number generator initialized with the integer 1933. For each interval, first draw the direction using randrange(4) in north-east-south-west order, then one uniform random() variate for the outcome, declaring wet if it is less than the sector probability. Initialize each slider at one and the preceding outcome to dry for

the persistence baseline. Intervals 1 through 200 served as an initial observation period, 201 through 600 as the stable test, and 601 through 1000 as the changed regime. All forecasts and scores were calculated before updating on the new outcome. These details are sufficient to reproduce the illustrative record with a compatible implementation of the generator. No numbers below claim to be field performance.

TABLE 2

Correct forecasts in a synthetic trial

Intervals	Trials	Four sliders	Always dry	Persistence
1–200	200	157	110	103
201–600	400	302	207	196
601–1000	400	294	199	178
601–620	20	11	8	7
621–1000	380	283	191	171

Each count gives the number of correct forecasts. Persistence predicts the previous interval's outcome. The last two rows break down the changed regime.

Two features of these results are worth inspecting. The predictor got 302 of the 400 stable test intervals right, while always reporting dry succeeded 207 times. The trial was explicitly constructed to reward the feature the sliders represent. It would have been surprising if no advantage appeared over a long enough period. Immediately after reversal, old memories worked against the learner. It scored 11 of the first 20 in the changed regime. In the remaining 380 it scored 283. A short run can mislead in either direction. Even the persistence baseline can be uncommonly lucky or unlucky in a finite sequence whose events are independent. We have demonstrated a response in a specified model world. The result is a reason to test a physical design and supplies predictions about its memory, not evidence of useful meteorology in a valley.

The long-run behavior under those artificial assumptions can be checked independently of the code. Consider just the intervals in one sector. Let p be its unchanging wet probability and q its dry probability, with q equal to one minus p. State transitions between neighboring positions occur upward with probability p and downward with probability q, with end stops holding any

outward step. In a stable distribution the probability of flow from state zero to one balances flow in the opposite direction. Thus the mass at state one relative to zero is p divided by q. Applying the same reasoning to the next two boundaries makes the four masses proportional to 1, r, r squared, and r cubed, where r is p divided by q. This derivation applies when both outcomes are possible and observations in that sector are independent.

The upper pair therefore carries probability r squared divided by one plus r squared. In terms of p and q, the chance of a wet forecast is p squared divided by the sum of p squared and q squared. The chance of being correct is the sum of p cubed and q cubed divided by that same denominator. At p equal to three quarters this gives 0.70. An oracle that knew the sector's true long-run probability and always predicted wet there would score 0.75. The learner pays for its ability to adapt from finite recent experience. At p equal to one half its expected score is 0.50. With no predictive regularity to exploit, no ingenuity in the cut of the wooden notches can conjure one from independent fair outcomes.

This result also explains why a finite success in the simulation is not the only relevant test. We have an independent theoretical prediction for a stationary model, a step-by-step trace for a deterministic test sequence, and a comparison of forecasts on a changing pseudo-random sequence. A mistake in programming can survive either the trace or the aggregate test but is less likely to survive agreement with all three. A mistake in the physical implementation can then be localized by comparing it with the already-tested abstract rule under controlled inputs. A contemporary investigator may use software to check an ancient-material thought experiment without sneaking a computer into its intended functioning.

What is a forecast worth to the village?

The four outcome bins from the synthetic stable test contain 151 correctly forecast wet intervals, 151 correctly forecast dry intervals, 56 false wet alarms, and 42 missed wet intervals. This particular equality between the two correct counts is an accident of the sequence. It lets us examine an application without inventing another dataset. Suppose the prediction is used to decide when a roof should be closed over drying goods. Closing has a cost c, including the lost drying opportunity and the effort of operating the roof. Leaving it open when it rains has a cost d in damage. For this small model assume closing completely

prevents the damage, that the loss c applies every time the roof is shut, and that its operation does not affect the separate rain gauge.

Closing whenever the device predicts wet would cost 207c for the 207 closures and 42d for the missed wet intervals. Leaving the roof open throughout would cost 193d. In the specified trial the weather rule beats leaving the roof open whenever 207c is less than 151d. Closing every interval would cost 400c. The forecast rule beats this precaution when 42d is less than 193c. Together the two inequalities describe costs for which the forecast beats both extreme policies. In approximate terms c divided by d must exceed 0.218 yet be less than 0.729. Even a statistically useful forecast may have no decisive practical advantage when protection is almost free or when closure is exceptionally burdensome.

This calculation also shows how easy it is to double-count an advantage. The 151 successful wet forecasts do not give 151 free savings: the cover had to close to realize each saving. The 56 false alarms do not merely decrease an accuracy score: they cause 56 real closures with no compensating avoided rain damage. The 42 missed wet intervals should not be charged once as prediction mistakes and again as a separate machinery failure in the same cost model. A device meant to serve a purpose needs an explicit connection between its categories and the outcomes relevant to that purpose. Numerical accuracy alone cannot tell a community what to build.

A more realistic assessment would ask whether the false alarms arrive on consecutive intervals. Closing and immediately reopening a roof may take more work than leaving it closed for three wet periods. Damage may saturate once goods have become thoroughly soaked, and drying opportunities may depend on temperature. Some missed intervals may matter much more than others. The narrow prediction problem can still be an entirely sound first experiment. An account that extrapolates directly from its score to food saved throughout the year has omitted an intervening model of action and consequence. That model can be tested later or replaced by a properly measured practical trial.

There is one final feature of the proposed experiment worth specifying in advance. A machine that fails on ten stormy intervals and therefore supplies only 390 predictions out of 400 can appear especially competent if one compares accuracy only where both the inputs and outputs happened to be legible. Failure to issue a forecast is itself a service result. A sensible record

would include the number of expected intervals, valid forecasts, unresolved wind observations, incomplete updates, and confirmed successful resets. A simple always-dry comparator can issue a prediction even when the vane is jammed. Fair comparison sometimes demands reporting both accuracy conditional on a working instrument and overall coverage across the full period. This is another place where a carefully designed set of mechanical witnesses can help a builder avoid writing a flattering story afterward.

None of these accounting choices calls for a human to decide during operation what a particular rainstorm meant. They concern how a future observer will evaluate the service and how the device can be arranged in advance to leave interpretable evidence. There is room for a substantial human act of judgment before the first interval and another after the last, with an autonomous series of committed forecasts and mechanically registered outcomes between them. That middle space is precisely the one our thought experiment set out to find.

Five years on the Kansas weather record

What a slider learns in a valley depends on what comes down the valley. We can give the rule some real weather. Parsons, in southeast Kansas, has an airport station whose routine hourly reports include a wind observation and a precipitation record. For six calendar years, 2019 through 2024, we retrieved these reports from the Iowa Environmental Mesonet archive. Imagine a flag being set at each eligible hourly reading. Its only input for choosing a slider is the direction from which the wind is blowing now. The question is whether the *following* hour will register at least one hundredth of an inch of precipitation. A trace falls below the threshold. The outcome is the station's reported accumulation, which can include precipitation other than rain. [34–36]

Every slider began in position one. A year of observations in 2019 supplied its initial history. That year also established the majority outcome for each season. Alongside the four wind memories we ran a single slider with the same four notches. Every eligible hour updated that one slider, whatever the direction of the wind. Three simpler rules called for dry weather, repeated the preceding hour, or used the 2019 majority for the season. The test ran from 2020 through 2024.

The thresholds, quadrant boundaries and update rules stayed fixed. Both learners kept their positions across gaps. Each forecast was recorded before its outcome could move a slider.

A qualifying pair consisted of two routine reports sixty minutes apart with usable precipitation at both ends and a numerical wind direction at the start. Calm or variable winds gave the vane no quadrant to select. The single slider could have issued a forecast in a calm, but every rule was scored on precisely the same hours. There were 5,780 eligible intervals for preparation in 2019 and 22,959 for testing in the next five years. The notes give the full selection rules. [34–36]

TABLE 3

Forecasts at Parsons, Kansas, 2020–2024

Year	Hours	Wet	Four sliders	One slider	Repeat	Dry / season
2020	3,611	225	3,383	3,418	3,461	3,386
2021	3,680	202	3,487	3,508	3,550	3,478
2022	5,595	216	5,379	5,393	5,441	5,379
2023	5,205	204	5,003	5,009	5,067	5,001
2024	4,868	212	4,626	4,646	4,710	4,656
Total	**22,959**	**1,059**	**21,878**	**21,974**	**22,229**	**21,900**

The last four columns count correct forecasts. Repeat uses the preceding hour. Dry was the 2019 majority in every season, so the dry and seasonal rules share one column.

A single slider made 21,974 correct forecasts. Routing observations to four memories brought that down to 21,878. The single slider was right in 258 hours when the four were wrong. The four were right in 162 hours when the single slider was wrong. That gives a net advantage of 96 hours to the simpler mechanism. The difference had the same sign in all five years. Repeating the preceding hour did better still, with 22,229 correct. A long dry spell makes every rule look good on this measure. What happened in the hours when someone might actually want a roof closed?

TABLE 4

What happened when a wet hour mattered

Forecast rule	Wet hits	False alarms	Wet misses	Correct dry	Accuracy
Four sliders	335	357	724	21,543	95.29%
One slider	459	385	600	21,515	95.71%
Repeat last hour	689	360	370	21,540	96.82%
Dry / seasonal	0	0	1,059	21,900	95.39%

A wet hit means measurable precipitation was forecast and observed. A false alarm is a wet forecast followed by an hour below the threshold.

The single slider warned of 459 of the 1,059 wet hours. Compared with the four sliders, that bought 124 extra warnings for 28 extra false alarms. Repeating the preceding hour did better on both counts than the single slider. Against the four sliders, repetition caught 354 more wet hours for only three more false alarms. A farmer can attach a cost to an unnecessary closure and a different cost to goods left in the rain. Overall accuracy hides those choices.

On May 4, 2024, one can actually watch the south slider lag behind a change. At 07:52 UTC it stood at zero and forecast dry. The hour ending at 08:52 was wet, raising the slider to one. The wind was still southerly. The slider again forecast dry, and the next hour was wet too. At 09:52 it finally occupied position two and forecast wet. The next hour fell below the measurable threshold. At 10:52 the wind belonged to the north quarter, and the machine consulted a different slider. These readings show what a saturating counter does with evidence that arrives in bursts. A two-hour run can pass before the flag powered by its first two results ever turns. One episode illustrates the mechanism. The five-year totals judge its performance.

Wet hours followed about 7.9 percent of easterly winds, 6.5 percent of northerlies, 4.1 percent of westerlies and 3.1 percent of southerlies. Those differences leave room for another rule to make better use of direction. For these sliders, splitting the rain history by quadrant meant that recent observations from other directions were lost to the next forecast. The simpler rival suggests a workshop experiment. The rain cup

could leave a single mark that lasts until the next tick. That mark would serve as the next forecast.

The 22,959 hours account for about 52.4 percent of the 43,848 clock hours in the five test years. Missing wind observations, calm or indeterminate directions, gaps, and precipitation-quality exclusions matter to what is being claimed. For instance, a rule that always says dry can run in dead calm, whereas this four-sector device has no input to latch. Restricting the score to matched opportunities gives a fair head-to-head comparison. It also means the reported percentages apply only when those opportunities occurred. The replay used one station, one lead time, and the published update rule. Its result belongs to that combination. It delivers a local answer about a forecasting rule while leaving the behavior of a wooden linkage exposed to wind and rain an open engineering question.

What the workshop would have to settle

One can investigate the instrument piecemeal without pretending its largest question is settled. Test a single counter made of appropriate materials while driving its two pawls through long sequences of alternating and repeated updates. Does state three remain at three after ten wet inputs? How often does a dry input fail to move a saturated slider one notch? Test the direction latch separately while a vane swings across boundaries during the update. Test the rain threshold while the timing vessel empties nearby. At each stage the trial has a specific physical claim. A device need not be exposed to a year of weather to discover that its cup never completely empties.

There is a trade between low friction and reliable commitment. A light lever follows a small wind but can flutter while sampling. A firm detent gives repeatability but demands enough force to overcome it. A wet cord changes length and can change the timing of a release. Wood swelling after rain can make a pawl especially likely to stick on the very intervals scored as wet. That last possibility is insidious. Mechanical error can correlate with the target being learned. A bench test in a dry shed may then certify a device that fails selectively in the field. The problem belongs to the design, not to a mysterious weakness of wooden intelligence.

A person would reasonably ask how much work is required just to preserve a dependable water supply. An intermittent spring can stop a timing cycle halfway through, turning a forecast for one interval into a gauge record for

several. A strictly independent machine can either invalidate interrupted intervals using a suitable mechanical condition, or be specified to operate only while the source stays within a range. A laboratory investigator may count invalidations after the fact when evaluating the instrument. If the published claim is that the device automatically judges its own performance, it cannot quietly depend on the investigator to rescue its score from interrupted runs. We have come full circle to the operator beside the hollows.

There are also useful choices about what to do with a forecast. The simplest flag merely displays it. One could couple the output to a cover over a small object, then score whether covering it saved exposure to rain at the cost of time unnecessarily covered. But operating the cover could affect the measurement if it shelters the gauge. A learner whose action hides the very outcome used to evaluate it encounters a much harder causal problem. For the first instrument, physical separation of the flag and the rain collector is an advantage. It makes forecasting testable before adding the effects of intervention. For a subsequent device meant to take useful action, those effects become a legitimate subject of their own.

Suppose the whole mechanism finally passes its bench tests and scores worse than persistence in a long field trial. That result would distinguish at least three questions. The system might be a functioning autonomous learner. Its chosen input, wind quarter at one instant, might be a weak predictor at that site and time scale. Its four-state update rule might discard useful structure even if such structure exists. A report that says only “Stone Age AI failed” would bury all three conclusions. A report of numerical success without an independent weather record would be little better. There is no shortcut around defining the problem whose solution is being claimed.

A builder at the river could, with less trouble, carve a traditional association straight into the flag linkage. It might even outperform the learner during a stable period, particularly if local observers had acquired excellent weather knowledge over years. The point of the harder design is that it can change its future response through events its maker never witnessed. If it begins in one valley and ends in another, the notches can come to stand for something different without being recut. The difference is small enough to point to in a drawing and important enough to give the book its title. After we walk away, which part of the next judgment can the world still write into our machine?

What Would an Archaeologist Find?

Suppose the last person who understands the weather machine leaves the valley. A few seasons later the spring changes course. The tipping vessel dries. A cord breaks, a vane falls, and the wettest piece of wood happens to survive longest. Centuries later someone finds four short slotted pieces and a stone basin. Beside them are a few possible pivots and evidence that water once ran close by. What kind of story could an honest archaeologist tell?

The case is deliberately unfair to the future investigator. We know the mechanism because we invented it. We know how hard one has to work to make its parts do anything at all. We can look at the hypothetical remains and mentally install the missing timing sequence without feeling how much evidence that sequence would require. The investigator who sees the objects for the first time has no such advantage. Perhaps the slots had a religious use. Perhaps they supported a sliding shutter, recorded the number of catches at a fishing station, or belonged to separate moments in a long occupation. Evidence that water once flowed beside them improves some stories and does fairly little for others. It does not uniquely restore the vanished order of operations.

One kind of find would be a piece joining two parts under dispute. If it fitted the latch and a directional selector, the unusual arrangement of holes could identify a possible coupling. Wear on only one side might favor one direction of movement. Matching groups at a second site could show that the arrangement was deliberate. Models could explore whether proposed parts would have moved with the energy available. A skeptical reconstruction might discover that a proposed heavy selector could never have lifted on the flow indicated by the channel. A beautiful explanation would then fail through an ordinary shortage of work. These are real avenues of inquiry even when the grandest interpretation stays uncertain.

What would distinguish a rain recorder from a rain predictor? A recorder might couple an observation directly to a mark. A predictor must commit to an outcome before the future observation that can judge it. A learner must also let the later outcome alter something used on a future occasion. These are distinctions about *when* a change occurs, *which earlier input* it is associated with, and *whether* the resulting state is used in making another prediction. An isolated notch preserves the possibility of a position. It seldom preserves the temporal clause. That is the reason an excavation could recover most of the weight of an intelligent device while recovering little of the evidence for its intelligence.

Conversely a trace that seems quite uninteresting could transform the argument. Imagine a series of marks associated with interchangeable forecast and outcome positions and an independently legible sequence of dates. The marks might reveal that a prediction was committed on an earlier date and that the device subsequently modified its response to a recurring cue. Several rival explanations would still need to be tested. The crucial evidence might be the relation across records rather than any single marvelous object. In our own time the same can be said of a dataset with recoverable provenance. The coordinates of an old model, the input it received, the output it committed to, and the later observation together establish something that no isolated eloquent answer ever could.

Imagine discovering a second copy whose directional memories are frozen in very different positions. There are at least three explanations. The devices could have learned under different local weather. A craftsperson might have set each to a traditional rule and pinned it in place. One machine might simply have failed earlier than the other, recording an unrepresentative series of events. The comparison gains force if there are independent records of the environment, different wear at the stops, or a mechanism that physically connects the rain cup to an addressed slider. Without some such evidence, calling the differences learning would be premature. Calling them meaningless decoration would be premature too. Good historical reasoning can support a live possibility without promoting it into a result.

The same discipline can be turned toward the traces contemporary civilization is deliberately preserving. A trained model's weights may be retained while the examples, preferences, and evaluation choices that gave them significance are lost. A successful robot's demonstration can survive as a

film after the code that generated the actions disappears. An electronic circuit may survive physically for decades while its programmable components lose their contents. A clay bowl can be more durable than the association it once held. So can a metal cabinet with an impressive inventory number. An artifact can look modern enough to be assumed completely documented long after its important operating history has gone missing.

If a future curator could keep only a small collection of things from one of our devices, I would want a design, a dated input, the committed response, the eventual outcome, and a second trial with a documented intervention. I would also want some record of the operator's instructions. That last item can look like ordinary office waste beside a piece of ingenious hardware. Without it nobody may know that an attendant was told to make a particular judgment after each run. Such records can prevent both overstatement and neglect. A future historian ought to be able to credit a machine with a genuine achievement that the contemporary audience did not notice, as readily as to discover the human behind a celebrated display.

What a civilization chooses to preserve is seldom decided by someone posing our question. An accountant may preserve a fuel bill that establishes a machine was operating at all. A repairer's note about a chronically sticking latch might supply the only independent description of the order in which parts moved. An editor's qualification about the path a prototype could actually follow can outlast the machine and protect its real achievement from a more romantic story. Ross gave the future reader that gift when he distinguished his planned third path from the paths achieved. An honest record of a small imperfection can tell us more about intelligence than the best surviving photograph of a perfect run.

The conversation that prompted this book had its own archaeological problem. People have long made things that carry knowledge and help produce answers. If a functional definition of intelligence can encompass some of that history, should we relabel the objects in our museums? Sometimes a more revealing label would help. One could describe how a valve senses a deviation or how a relay acquires a preference, rather than merely giving the date, manufacturer, and materials. A title such as “Mesolithic artificial intelligence” on a fish trap would move in the other direction. It would give a disputable conclusion just where readers need an account of the operation. Preserving differences among unattended action, regulation, learning, and

inquiry makes it possible to recognize genuine achievements even when their makers spoke a different technical language.

The history also bears on a more distant sort of investigation. Suppose observations of a planet suggest that some process maintains a condition long after an initial disturbance. We might be tempted to speak of regulation or purpose. Our visit to the float teaches us to look for a causal path by which departure from a condition changes the response. Suppose an apparent response to a recurring event occurs earlier on later occasions. Ross's arm teaches us to look for a trace of the first encounter that could alter the later one. Suppose a series of changes systematically distinguishes among rival possibilities. A robot laboratory teaches us that such a pattern would call for evidence about which observations selected which interventions. These are ways to frame questions about an unknown system, not grounds for claiming any particular planet has passed the tests.

The same habits matter in assessing very familiar machines. A conversational AI can be impressive at precisely those moments when independent checking is cheapest to omit. A spacecraft that acts competently within its authority can be impressive at moments when nobody except an engineer looking at a log would notice it. A laboratory instrument can be most useful when it designs a trial whose outcome disappoints the person who commissioned it. A headline about machine intelligence tends to carry the whole achievement away in one phrase. Following the individual signals allows some of the credit to return to a questioner, a craftsperson, a circuit, an earlier investigator, or that most refractory participant in scientific work, the object under examination.

There is a temptation to end such a history with an escalating prediction. If two currents of water and four sliders can learn, a sufficiently grand machine must eventually do everything we value in intelligence. The inference lacks a premise. A bigger device can repeat a small mistake at greater scale. A more flexible one can solve a problem its designer never considered, while creating the need for forms of accountability its predecessor never required. Universality supplies important mathematical common ground for machines of astonishingly different kinds. It promises neither the data on which a good judgment depends nor the wisdom to decide which judgments a community should delegate.

A better end is an experiment imaginable at the beginning. A builder finishes an apparatus and makes a record of what was put into it. The initial wind is from the east. The flag says dry. Rain arrives during several intervals of the relevant kind, some when nobody is present. Weeks later the same initial wind sets the flag to wet. An investigator returns and finds the east slider near its wet end. If the mechanism held its forecast in advance, associated later outcomes with the right observations, and used the resulting memory on the later occasion, something quite definite has been learned. Whether it is a good forecast in that valley is still up to the next rains.

The community could reasonably take pride in making an object in which such a thing can happen. It is useful to say exactly where the pride belongs. The builders made it possible for experience to leave a consequence. They had no opportunity to insert the particular rain that fell after their departure. The world contributed that. The machine's contribution was to let an earlier encounter with that world alter what it did when the wind came round again.

Notes on Sources and Experiments

Reading an engineering claim

There are several different sorts of evidence in this history. A surviving object can establish materials and some of the physical relations among them. An inventor's report can establish the intended sequence of a trial and distinguish configurations that a photograph alone leaves ambiguous. An independent reproduction can test how well the sequence works within a reconstructed setting. A modern scholar can connect a claim with a date, an earlier description, or a document in a different language. These sources have different strengths. The numbered references identify the principal support for historical and technical claims. These notes also describe the two original simulations and the replay of hourly weather observations from Parsons, Kansas.

The prologue and Chapters 1–4

Ross's published diagrams are the basis for the opening description [1]. His second path shows residual entries into blind passages. The third, more abbreviated path illustrates possible improvement. Descriptions of changing a single switch or controlling for friction are proposed tests. They make no claim to be tests documented in the original paper. The stronger case for a particular apparatus often comes from an author who carefully marks the boundary of the achieved result.

The Liffey account is based on the excavators' paper, published in the 2007 print volume of Antiquity [2]. The field dates are calibrated prehistoric dates. They are not estimates of the age of every surviving wooden trap in northern Europe. The generic funnel and successive hypothetical gates introduced in Chapter 1 serve to distinguish different distributions of human and mechanical work. The Marshallese comparison concerns how knowledge is carried during an activity, with no chronological identification of the two traditions [27].

Historical reports about Ctesibius are transmitted through later writers. The mechanism described by Mayr is the basis for the float-regulation example [3]. The Banū Mūsā lamp belongs to a medieval textual design tradition [4]. The modern mathematical examples about leaks and delays explore principles of control. No rates or dimensions in those examples are offered as measurements on either historical instrument. The 1788 engine, in contrast, has a surviving museum identity [5]. Shepherd discusses the fantail as part of the historical development of windmill control [28].

Michie's matchbox machine was built for noughts and crosses [6]. The seven-stone game and its independent simulation are original illustrations. The digital experiment used Python's standard Random generator with integer seeds 1933 through 2052 inclusive, separately for each training condition. Draws used integer selection over the total marker count, with take-one markers occupying the first interval. The uniform training opponent used equal probability over legal moves. Moves used no random draw for a deterministic opponent. A winning episode increased every marker count corresponding to a selected move by one. A losing episode decreased each selected count by one, with a floor of one. The ground truth of a terminal position was assigned to the player who had just moved. Frozen-policy evaluation used exact dynamic programming. Rounded table entries describe mean probabilities. Results would change under other reinforcement conventions or alternative reliable opponents.

The chess position in Chapter 4 is an example for the reader. It should not be treated as a position recorded in a historical demonstration. Marsh distinguishes Torres Quevedo's earlier upright machine, later horizontal version, special starting conditions, and restrictions of the endgame [7]. The anecdote about consulting the zā’irja on its own antiquity is discussed by historians of Ibn Khaldūn's engagement with the device [29]. It is evidence for the history of a practice, not independent verification of the information returned by that practice.

Chapters 5–9

Ross and Shannon offer two different opportunities to read the builder's account of a demonstration [1, 8]. Ross uses the language of conditioning. Shannon explains relays that retain a last direction and a change between exploration and use of remembered directions. His presentation includes questions from an audience watching the apparatus. The argument from increasing times of last departure in Chapter 5 is a mathematical explanation for the reader, applicable to a completed walk ending at its goal in a static graph. The invented lettered

walk is separate from Shannon's historical example. The ant comparison draws on research about trail reinforcement in particular species and environments [9].

Ashby's account specifies the selectable parameters and the distinction between different levels of feedback [10]. The independent-draw search calculation is deliberately simpler than the homeostat. It is an example of how the distribution of acceptable arrangements and the time permitted can matter more than an impressive count of possible combinations. No search durations from that numerical example are measurements on Ashby's apparatus.

Samuel's paper should be read for both the ingenuity of the method and its discussion of its own limitations [11]. It documents different learning schemes and qualifications concerning some of the comparisons. The checkers evaluation with only material and mobility and the four-corner threshold exercise are teaching examples. The optical and motor-driven details of the Mark I are historical [12]. For the variable-resistor network, the pertinent evidence is the iris test set and the local learning scheme in the original paper [13]. The test involved classifying presented flower measurements. It did not include the independent botanical measurement of intact plants.

The rule about worrying and the example of two wells were written for Chapter 8. Weizenbaum explains the actual historical program's use of decomposition and reassembly rules [14]. The cart-and-door example similarly illustrates dependencies in planning, while the evidence for Shakey as a working integration of perception, planning, and action comes from SRI [15]. For Deep Space 1, the NASA release gives an accessible chronology, and the technical report distinguishes components and discusses validation [16, 17]. The ten-unit power budget is an invented illustration. The 1999 flight result was an experiment with bounded periods of onboard control.

Chapters 10–15 and epilogue

Read the Adam paper's evidence table as well as its abstract [18]. An inference from growth, a subsequent manual test of biochemical activity, and agreement with previous literature do different work in the historical claim. The cancellations, instrument-position effects, and three-explanation exercises are examples of experimental reasoning independent of the reported hardware. The Go account concerns AlphaGo Zero as described in 2017 [19]. It should not be used to infer what training data went into every system in the wider AlphaGo family.

The transformer, GPT-3 and ChatGPT launch accounts document the historical sequence [20–22]. The marked machine-voice passage describes the exchange that led to the book. Pages 98–99 give the prompts, replies and scoring rule for a four-case demonstration during revision. The assistant had the chapter in context. The proposed tests with a fresh session and independent scorer remain for a future trial.

The two-counter construction uses the conventional idealization of expandable nonnegative counters and a finite controller [23, 24]. Its equivalence to a tape model presumes suitable encodings. The account of physical hypercomputation distinguishes a mathematical limit on effective procedures from the further question of which physical processes can actually be realized [25].

The four-state wind predictor has the same update form as a two-bit saturating counter [26]. Its weather application, mechanical organization, mathematical calculation, and illustrated synthetic series were developed for this book. The input probabilities in Chapter 15 were chosen by construction, and the synthetic record represents no location on Earth. For the stable test, the cost example assumes full protection when closed and a known flat cost for each action, independently of previous actions. The first twenty changed-regime intervals form a subset of the four hundred changed-regime intervals. Adding their rows as if they were separate episodes would double-count them. The imagined remains in the epilogue refer to that proposed device alone.

Additional evidence and the Parsons replay

The visible outputs of Zhang Heng's instrument are described in historical accounts and in the Science Museum Group record [30]. Yan and Hsiao explain how the surviving description admits multiple mechanical implementations [31]. The comparison with calibration and repeat-event capacity is an engineering inference. For the slime mould, the starting configuration, the two pairs of path alternatives, and the mixed outcomes come from the 2000 report [32]. In Chapter 13 the three-cord principle follows Park and Flores [33]. The right-triangle dimensions and path lengths were chosen and calculated for this book. The proposed fiber-and-stone version has no claimed archaeological provenance.

To reproduce the weather trial, retrieve the routine hourly reports for station PPF in the KS_ASOS network from the Iowa Environmental Mesonet, one UTC calendar year at a time, 2019–2024 inclusive [34, 35]. Request the fields valid time, wind direction in degrees, wind speed in knots, one-hour precipitation in

inches, and raw METAR text. The corresponding field names are valid, drct, sknt, p01i, and metar. Use observation type 3 (routine), a distinct trace marker T, and a distinct missing marker M. The archive end date is exclusive, so the final request ends January 1, 2025. The trials used the archive as retrieved on November 18, 2026.

Sort reports by UTC observation time. Use only adjacent timestamps with exactly one report at each, both at minute 52, exactly one hour apart. Require the AO2 station remark at both endpoints, usable precipitation amounts at both ends, and no PNO or missing hourly precipitation remark at either end. The ASOS manual explains that P0000 denotes a trace and that an otherwise functioning gauge can omit the hourly precipitation group when the amount is zero [36]. A reported trace or amount below 0.01 inch is dry. Reject ambiguous, missing, or calm initial wind. Require positive speed, a numerical initial direction, and a directional wind group in the original METAR. Turning through north, the quarters begin at 315 degrees for north, 45 for east, 135 for south, and 225 for west, each including its lower boundary. A 360-degree wind is north.

At each eligible starting time use the position of the selected quarter before any update from the following hour. Positions zero and one predict dry. Two and three predict wet. After observing the outcome, increase the position by one if wet and decrease it by one if dry, stopping at three or zero. Skipped pairs make no update. Retain the four positions across gaps and year boundaries. For persistence, take the wet or dry observation already available at the starting report. For the seasonal baseline use December–February, March–May, June–August, and September–November by the starting UTC month. Count the wet and dry outcomes of the eligible 2019 calibration pairs in each season and take its majority, breaking any tie toward dry. Assign each scored outcome to the UTC year of the ending report.

An independent replay decoded the numerical wind and hourly precipitation groups directly from the archived METAR strings. It matched the selected times, quarters, observed outcomes, and forecasts of the replay using the archive's decoded fields. The archive occasionally contains unavailable observations. Leaving both the numerator and denominator visible makes the distinction between success on an eligible hour and service throughout an entire year explicit.

For the one-slider control, initialize one counter at one before the 2019 preparation period. Predict wet from positions two and three, then update that same counter after every eligible outcome. Carry its state across gaps and year boundaries. The revised replay reproduced the counts for the original rules. Among the test hours the two slider forecasts disagreed 420 times.

References

[1] Thomas Ross. “Machines That Think.” Scientific American 148, no. 4 (April 1933), 206–209. DOI 10.1038/scientificamerican0433-206. See particularly Figures 2–6 and the discussion of the different paths the sensory tip traced.

[2] Melanie McQuade and Lorna O’Donnell. “Late Mesolithic Fish Traps from the Liffey Estuary, Dublin, Ireland.” Antiquity 81, no. 313 (2007), 569–584. DOI 10.1017/S0003598X00095594.

[3] Otto Mayr. Feedback Mechanisms in the Historical Collections of the National Museum of History and Technology. Smithsonian Studies in History and Technology, no. 12, 1971. See the discussion of early water-level regulation and Figure 2.

[4] Banū Mūsā ibn Shākir. The Book of Ingenious Devices. Translated and annotated by Donald R. Hill. D. Reidel, 1979. For the lamp mechanism, see also the Smithsonian Lemelson Center, “Ingenious Devices,” discussion of model 96.

[5] Science Museum Group. “Rotative Steam Engine by Boulton and Watt, 1788.” Collection object co50948. For an account of the regulating action, see also Mayr [3].

[6] Donald Michie. “Experiments on the Mechanization of Game-Learning. Part I. Characterization of the Model and Its Parameters.” The Computer Journal 6, no. 3 (1963), 232–236. DOI 10.1093/comjnl/6.3.232.

[7] Allison Marsh. “This 1920 Chess Automaton Was Wired to Win.” IEEE Spectrum, 2023. The article distinguishes the two versions of El Ajedrecista and describes the limits of the demonstration. spectrum.ieee.org/chess-computer-1920

[8] Claude E. Shannon. “Presentation of a Maze-Solving Machine.” Transactions of the Eighth Macy Conference on Cybernetics, published 1952. Reprinted in N. J. A. Sloane and Aaron D. Wyner, eds., Claude Elwood Shannon: Collected Papers (IEEE Press, 1993). See the audience discussion and Shannon's account of relay memory.

[9] Raquel G. Loreto et al. “Foraging Ants Trade Off Further for Faster: Use of Natural Bridges and Trunk Trail Permanency in Carpenter Ants.” Naturwissenschaften 100 (2013), 957–963. DOI 10.1007/s00114-013-1096-4.

[10] W. Ross Ashby. Design for a Brain: The Origin of Adaptive Behaviour. Chapman & Hall, 1952. See the chapter on the homeostat, including the roles of the stepping switches. The W. Ross Ashby Digital Archive preserves the book and his journals. ashby.info

[11] Arthur L. Samuel. “Some Studies in Machine Learning Using the Game of Checkers.” IBM Journal of Research and Development 3, no. 3 (1959), 210–229. DOI 10.1147/rd.33.0210.

[12] National Museum of American History. “Electronic Neural Network, Mark I Perceptron.” Smithsonian collection object nmah_334414. Nils J. Nilsson, The Quest for Artificial Intelligence (Cambridge University Press, 2010), discusses the public demonstration and the electrical implementation of adjustable weights.

[13] Sam Dillavou, Menachem Stern, Andrea J. Liu, and Douglas J. Durian. “Demonstration of Decentralized, Physics-Driven Learning.” Physical Review Applied 18 (2022), 014040. DOI 10.1103/PhysRevApplied.18.014040. See Figure 3 and the task-specific experimental details.

[14] Joseph Weizenbaum. “ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine.” Communications of the ACM 9, no. 1 (1966), 36–45. DOI 10.1145/365153.365168.

[15] SRI International. “Shakey the Robot.” Artificial Intelligence Center research, 1966–1972. SRI's technical overview and the Computer History Museum's Shakey exhibit describe the experimental environment and capabilities. sri.com/hoi/shakey-the-robot/

[16] NASA Jet Propulsion Laboratory. “Remote Agent Experiment Meets All Objectives.” June 7, 1999. The release distinguishes the simulated camera, electronics, sensor, and thruster faults from the actual timing problem encountered during the flight test.

[17] P. Pandurang Nayak et al. “Validating the DS1 Remote Agent Experiment.” International Symposium on Artificial Intelligence, Robotics and Automation in Space, 1999. The flight report describes the goal-driven architecture, validation process, and on-orbit experiments.

[18] Ross D. King et al. “The Automation of Science.” Science 324, no. 5923 (2009), 85–89. DOI 10.1126/science.1165620. The main text and supporting material qualify the experimental evidence and describe human follow-up.

[19] David Silver et al. “Mastering the Game of Go without Human Knowledge.” Nature 550 (2017), 354–359. DOI 10.1038/nature24270. The described system is AlphaGo Zero.

[20] Ashish Vaswani et al. “Attention Is All You Need.” Advances in Neural Information Processing Systems 30 (2017). arXiv 1706.03762.

[21] Tom B. Brown et al. “Language Models Are Few-Shot Learners.” Advances in Neural Information Processing Systems 33 (2020). arXiv 2005.14165.

[22] OpenAI. “Introducing ChatGPT.” November 30, 2022. The contemporary launch account describes dialogue training and the role of human feedback. openai.com/index/chatgpt/

[23] Alan M. Turing. “On Computable Numbers, with an Application to the Entscheidungsproblem.” Proceedings of the London Mathematical Society, second series, 42 (1936–1937), 230–265. DOI 10.1112/plms/s2-42.1.230. See also “Turing Machines,” Stanford Encyclopedia of Philosophy, for modern definitions.

[24] Cornell University. CS 4820, “Notes on Turing Machines.” See the simulation of stacks, counter automata, and the encoding of several counters by prime powers. cs.cornell.edu/courses/cs4820/2024sp/notes/turingm.pdf

[25] Stanford Encyclopedia of Philosophy. “The Church-Turing Thesis.” See the discussion of computability, stronger physical claims, and hypercomputation. plato.stanford.edu/entries/church-turing/

[26] Cornell University. ECE 4750, Computer Architecture. “Advanced Processors: Branch Prediction,” Topic 10 problems. See the four-state saturating counter. The forecasting application and analysis in Chapter 15 are independent of processor architecture.

[27] Library of Congress, Geography and Map Division. “The Unique Seafaring Charts of the Marshall Islands.” Worlds Revealed, November 2021. See also the Smithsonian Ocean account, “Navigating the Waters with Micronesian Stick Charts.”

[28] Dennis G. Shepherd. “Historical Development of the Windmill.” In Wind Turbine Technology: Fundamental Concepts of Wind Turbine Engineering, edited by David A. Spera. See the account of Edmund Lee's fantail and automatic orientation of the mill.

[29] Ibn Khaldūn. The Muqaddimah, translated by Franz Rosenthal. For the North African demonstration and its historical setting, see Robert Irwin, “Al-Maqrīzī and Ibn Khaldūn, Historians of the Unseen,” Mamlūk Studies Review 7, no. 2 (2003), 217–230.

[30] Science Museum Group. “Model of the Zhang Heng Seismoscope.” Collection object 1976-107. The cutaway illustrates one proposed interior. collection.sciencemuseumgroup.org.uk/objects/co54055

[31] Hong-Sen Yan and Kuo-Hung Hsiao. “Reconstruction Design of the Lost Seismoscope of Ancient China.” Mechanism and Machine Theory 42, no. 12 (2007), 1601–1617. DOI 10.1016/j.mechmachtheory.2007.01.003.

[32] Toshiyuki Nakagaki, Hiroyasu Yamada, and Ágota Tóth. “Maze-Solving by an Amoeboid Organism.” Nature 407 (2000), 470. DOI 10.1038/35035159. See Figure 1d for the frequencies of alternative outcomes. The same authors describe tube reorganization in “Path Finding by Tube Morphogenesis in an Amoeboid Organism,” Biophysical Chemistry 92 (2001), 47–52. DOI 10.1016/S0301-4622(01)00179-X.

[33] Jungeun Park and Alfinio Flores. “Fermat's Point from Five Perspectives.” International Journal of Mathematical Education in Science and Technology 46, no. 3 (2015), 425–441. DOI 10.1080/0020739X.2014.979894. The cord instrument appears in section 2.

[34] Iowa State University, Iowa Environmental Mesonet. ASOS/AWOS/METAR Data Download. Kansas airport network KS_ASOS, station PPF, routine hourly reports, January 1, 2019 through December 31, 2024, UTC. Retrieved November 18, 2026. mesonet.agron.iastate.edu/request/download.phtml?network=KS_ASOS

[35] Iowa Environmental Mesonet. Station Information for PPF, Parsons, Kansas. Network KS_ASOS. mesonet.agron.iastate.edu/sites/site.php?station=PPF&network=KS_ASOS

[36] National Oceanic and Atmospheric Administration and collaborating agencies. Automated Surface Observing System (ASOS) User's Guide (1998). See the hourly precipitation remark and the missing-precipitation indicator.