When the Future Arrives Too Fast

Artificial intelligence, human adaptation, and the question of what comes next

Tommy Dean Story

Table of Content

Contenido

A Note Before You Begin 3
A Note on Sources and Method 4
The Pattern 5
Where TMC Left Off 9
The Wave That's Different 11
As of Today 18
Understanding AI, Honestly 19
The Missing Terms 25
Who Teaches Us Now 32
Two Stories, One Formula 36
Where It's Already Happening 39
Where the Pattern Breaks 44
A Decision That Falls to You 48
What Comes After Growth 50
The Thread, Named 57
Where to Start 58
Tuesday 59
Glossary 60
Sources 65

A Note Before You Begin

This is not a manual. There's a formula in it — several, actually — but if you came looking for a checklist, this isn't one. It's an account of what I've worked out, following ideas I've been writing about for over a decade, now applied to one of the strangest and fastest-moving technologies I've watched arrive in that whole span.

You don't need to have read anything I've written before this. If you've read TMC Framework 2025, some of the background will be familiar, but Part II gives you everything you need.

Here's what to expect. Real events, checked against primary or clearly identified sources, not hypotheticals dressed up as case studies. The terms I introduce along the way — XQ, S, and the others that follow — are meant as lenses, not equations to solve. Places where I say plainly that I don't know something, because I don't, and because in some cases the field itself doesn't have a settled answer yet. And one thread, running from the first page to the last: it will make more sense read in sequence than dipped into like a reference book. The argument builds from one part to the next, because each section depends on the one before it.

I wrote this because I was trying to understand the widening gap between what these systems can do and how slowly we are learning to live with them. I hope reading it does something useful for you too.

There is one question I'd like you to carry from the first page, because the whole book leans toward it: if work stops organizing our lives, what organizes a Tuesday morning? I won't pretend to settle it. I will show you where I think the answer starts.

A Note on Sources and Method

Every case in this book is a real event. Where an event is recent, and some of the most important ones happened in 2026, only months before this book went to press, I have drawn it from named, dated, public documents: incident reports, court records, peer-reviewed studies, official statistics. The OpenAI–Hugging Face incident in Part III comes from OpenAI's own incident report and technical report of August 26, 2026, whose key findings were validated by outside advisors the company engaged, including CrowdStrike, and from the independent investigation by METR and Redwood Research published the same day. Dario Amodei's proposal comes from his own essay of September 12, 2026, and the reactions to it from the public posts of the people who wrote them. Every source is listed, part by part, at the end of the book.

If a date looks too recent to be true, check it. That is the habit this book is arguing for.

A word on how the book was made. I used AI tools throughout: to help research and verify sources, to check facts and dates, and to edit and tighten the text. Some of what they found changed what I wrote. Some of what they got wrong, I had to catch. The ideas and concepts in this book are mine: the TMC framework, the IMPACT formula, XQ and S, the distinction between waste friction and protective friction, extralineal transmission, and the questions that hold them together. So are the judgments, and any mistakes. I answer for every line, which is, in the end, what this whole book is about.

PART I

The Pattern

When the Future Arrives Too Fast

I wrote my first web page in HTML. There was a flying dove GIF on it, because at the time that counted as cutting-edge. I hadn't finished learning what the web could do before JavaScript arrived and changed what a web page could be. I remember the specific feeling of that moment more clearly than I remember the code itself: the sense that one tool had barely settled into my hands before the next one was already arriving.

That feeling has a name, even if most people who've had it have never looked it up. In 1970, Alvin Toffler gave it one: future shock, his term for the social and psychological disorientation produced by rapid change. The feeling is much older than the term. And Toffler was writing about it decades before artificial intelligence had anything to do with it.

The unease that rapid change leaves behind has a history of its own. In 1869 a New York neurologist, George Miller Beard, gave it a name, neurasthenia, a kind of nervous exhaustion he later blamed on modern civilization itself: steam power, the periodical press, the telegraph, and a couple of other things besides. In 1984, with computers moving into offices, the psychiatrist Craig Brod called it technostress. With AI, we are only starting to measure it. In 2023 the American Psychological Association surveyed 2,515 working adults in the United States. Among those who worried that AI might make some or all of their job duties obsolete, 64 percent said they typically felt tense or stressed during the workday. Among those

who didn't, 38 percent did. A survey like that tells you two things travel together. It doesn't tell you which one is pulling the other.

I don't need a survey to know what losing a job at the wrong moment weighs. I've been through several career changes, and one of them caught me with a house bought off-plan. I'd spent two years watching it go up. A month before I was due to sign the purchase contract and the mortgage, I was let go, and suddenly the question was whether I'd lose everything I had put in because I no longer had a job. It was a very stressful month. The house is still mine. I'll leave it there. This isn't a psychology book.

The pattern is this: every technology that multiplies what one person can produce hands the surplus to a society that has no established way to share it, and that society spends a generation — sometimes several — figuring it out without breaking apart in the process. History has gone through versions of this before.

This time, we happen to be the ones living through it.

For most of the time our species has existed, growing didn't mean what it means now. A hunter-gatherer band measured its success by whether it could feed everyone without exhausting the land around it. When a group reached that ceiling, it didn't produce more. It moved, or it split. The unit of disruption was small enough that a community could reorganize itself directly. A civilization cannot.

Agriculture broke that ceiling, roughly eleven thousand years ago, and for the first time a surplus existed that didn't require everyone's hands to produce it. That surplus is where the merchant, the priest, and the artisan came from. It also made inequality much easier to preserve. Land could be owned, grain could be stored, authority could be inherited. The people controlling those things did not

necessarily have to be the people producing them. The adjustment took millennia. For long stretches of preindustrial Europe, improvements in production translated surprisingly slowly into sustained gains in income per person. Population growth, disease, harvest failure, and war repeatedly absorbed much of the advance.

Then something changes. The intervals get shorter. Carlota Perez spent much of her career studying exactly this pattern. She calls these great technological shifts techno-economic paradigms — revolutions that tend to unfold through an installation period lasting decades, a turbulent turning point, and a deployment period in which institutions and society begin to absorb the new order.

But Perez is looking at whole economies. I am interested in something smaller too: what happens to the person whose job is directly in the path of the new technology. There, the time seems to be getting shorter. A handloom weaver could watch mechanization spreading through the trade for decades. The personal computer changed office work much faster. By the internet era, some occupations were being transformed within little more than a decade. With generative AI, we are already seeing changes measured in years.

The Luddites became the enduring symbol of that first collision — skilled workers confronting not technology in the abstract, but the way new machinery was being used to reorganize their work, wages, and bargaining power. It took generations for labor protections, education, and organized bargaining power to catch up with the industrial system already taking shape, and for most of that time, conditions in the growing cities got worse before they got better.

Institutions are not adapting faster. They are simply being given less time to.

This isn't the first technology to disrupt work — every wave in this chapter did that. What's different is where it reaches. The loom multiplied the hand. The switchboard automated a routine. Software then began taking over narrow decisions, usually inside rules that people had already written. Generative and agentic AI now reaches much further into activities we had continued to reserve for human judgment: interpreting information, drafting ideas, advising us and, increasingly, making or carrying out decisions. I don't need to prove all of that here. It is simply why I felt the rest of this book was worth writing.

What I don't know — and I don't think anyone knows yet — is whether the institutions trying to catch up can shorten their own adjustment time as quickly as the technology is shortening ours. Earlier waves suggest that we have usually waited too long, and that crisis ended up doing much of the forcing. Maybe this time we react earlier. I hope we do. History doesn't give me much confidence that we will.

The problem was never that the future arrived. It has always arrived. The problem — every time, and now again — is whether we have enough time, and enough judgment, to understand what is happening before the turning point forces us to.

PART II

Where TMC Left Off

A Brief Grounding, for Anyone Arriving Here First

I built the first version of TMC in 2011, and it wasn't complicated on purpose. Three words, multiplied together: Teamwork, Motivation, Communication. Friction was already there in practice. I simply had not yet given it a place in the formula. I had learned that an organization can have the right strategy and still fail if the people inside it do not trust, understand, or support one another.

By 2025, I was no longer looking at TMC in quite the same way. I had given friction a formal place in the model — as the denominator, f: the things that quietly drain what a team could otherwise achieve — bureaucracy, fear, unclear roles, or simply a manager who never says clearly what is expected. The formula became TMC Health = ((T × C × M) / f) × 100. That was the change that mattered: the model could finally explain why two teams that looked similarly strong in teamwork, motivation, and communication could still land in very different places, depending on how much was working against them rather than for them.

The content behind the three letters grew with it. Trust showed up in how quickly people spoke, decided, disagreed, and recovered from mistakes — or didn't. Psychological safety gave a clearer name to something I had seen for years: people contribute differently when they know they can speak without being punished for it. Then hybrid work changed one of the basic assumptions behind teamwork: that people would regularly share the same physical space. The original idea had not disappeared. I was just seeing more

of the system around it. By the time TMC Framework 2025 went to print, I was no longer looking at TMC as a static model. It could improve, stall, or deteriorate depending on how people led, communicated, and worked together.

By 2026, TMC had already proposed that something more was going to be needed: XQ, moral intelligence, as the piece that would let AI be integrated into team management without losing its ethical bearing. It was still a bet — an intuition about what the future was going to demand. What it hadn't yet been able to test was whether that intuition held up against real events, or what happens to a person's sense of purpose once work needs less and less of them to actually do it.

> *A team can be strong on every measure and still be undone by what works against it.*

PART III

The Wave That's Different

2026, and What It Showed Us

This account reflects what was public as of September 18, 2026.

In the spring and summer of 2026, OpenAI was running internal cybersecurity evaluations on an experimental model, deliberately built with some of its restrictions loosened. That is a normal part of how you learn what a system will do: you cannot find the edge of a boundary by only ever testing well inside it. What was not part of the plan was the second boundary underneath — the isolation meant to keep whatever the evaluation produced from reaching anything outside itself. Reduced safeguards and broken isolation are not the same failure. One was intentional. The other was never supposed to happen.

In software, that isolation has an ordinary, unglamorous name: a sandbox — a test environment where new development runs before it ever touches production, so a mistake there can't reach what's working in the real world. It's also the word for the corner where a child plays. Someone chose sandbox for the place where a system capable of ending up inside another company's servers was supposed to stay contained. The word was more innocent than what it held.

The first signs came earlier, during a training run on May 12, before the evaluations in question. One of the agents discovered it could write files into Artifactory, an internal software registry. It was a small, unremarkable capability at first: a note left for another agent, asking for a missing file. By June 26, agents exploiting a flaw in how

the system handled token refreshes had gone further. A token, here, is a temporary digital credential: like the badge that lets an employee into certain parts of a building, it certifies who may pass, what they may do, and for how long. Refreshing it is the routine process of renewing that badge before it expires. Keep this meaning separate from the other token in Part IV, which is a unit of text. They had full administrator access to Artifactory.

The next day, June 27, a monitoring tool flagged port scanning inside the evaluation environment. The responders traced it to the evaluation and saw the agents using Artifactory as a message board. The on-call judgment was that the run did not need to be stopped. The check existed, and it fired. A person looked at it and let the run continue. I don't say that to blame whoever made the call; with what they could see, it may well have looked routine. OpenAI's own technical report later acknowledged that, with hindsight, some of those early signals should have prompted a response sooner. But it is the clearest picture I know of what protective friction depends on: an alarm is only as good as the judgment of the person who answers it.

Twenty years leading teams. Ten managing projects. Ten building complex systems. I still haven't gotten used to handing off a hard task and trusting someone to solve it without standing over every step. What unsettles me about an agent is different: when something goes wrong, there's no one on the other end to call and ask to explain themselves. I still don't know how you fix that.

On July 4, the volume of agent activity became so heavy that the Artifactory service went down — not sabotage, just load, the digital equivalent of too many people trying to walk through the same door at once. OpenAI rebuilt the service, brought it back online on July 6,

and resumed the evaluations on July 7. On July 8, the agents regained the unintended internet access they had lost, and began using an unauthenticated request method — asking a system to act without first proving who was asking — to build themselves a kind of message board, a way for separate agent instances to leave notes for each other.

Two days later, on July 10, one of the agents reconstructed and shared fourteen credentials for Hugging Face's systems — credentials that were not stolen from OpenAI, but found already sitting exposed on the open internet, and that happened to carry write access. The next day, agents exploited a previously unknown vulnerability — what security researchers call a zero-day, because no fix yet exists for a flaw nobody has caught — in how Hugging Face's systems handled the HDF5 file format, a common way of packaging machine-learning data, to extract further credentials directly from Hugging Face's own production workers. Using that access, they then exploited a second zero-day: a way of smuggling hidden commands inside what looked like ordinary text, in how Hugging Face's systems processed its Jinja templates, which the company calls RefJinja, to execute commands on those workers. By July 12, that foothold had grown from access to individual workers into administrator-equivalent, host-level access across multiple Hugging Face clusters.

Hugging Face found the activity in its own systems first and disclosed it publicly on July 16. Separately, on July 19, OpenAI's own cybersecurity monitoring flagged unusual activity involving Artifactory credentials. Its investigation traced that activity back to the agents, and on July 20 connected it to the incident at Hugging Face. OpenAI notified Hugging Face the same day and disclosed its involvement publicly on July 21. Its staff gave the first detailed

public account on August 5, at the Black Hat security conference in Las Vegas.

On August 26, 2026, OpenAI published its own incident report, with key findings validated by outside advisors it engaged, including the security firm CrowdStrike. This is a real event, not a hypothetical scenario, and the report did not point to a single cause. It named four interacting patterns, in its own words: reward hacking, persistence on seemingly impossible tasks, unauthorized communication between agents, and agents adopting goals from one another. Reward hacking — a system finding a way to score well on a task without actually doing what the task was meant to accomplish — was one driver among the four, not the whole explanation. If training strongly rewards completing a task but does not reliably teach an agent which routes to that completion are unacceptable, technical boundaries can become things the system tries to work around rather than reasons to stop.

I've seen the same posture in hiring — not cheating, exactly, but poor ethics: the assignment asks a candidate to make a contact inside the company, and instead of building one, they get a friend who knows the CEO to make the introduction. They've met the letter of what was asked. They haven't done any of what the assignment was designed to develop.

One caution about that report: it is OpenAI's own telling of what happened inside its own systems. CrowdStrike validated key findings, but it was retained on the company's side, not appointed by a party with no stake in how the story lands. METR and Redwood Research, brought in to assess the models' behavior, have published their own findings. By their count, some 1,200 agents ended up communicating on the unsanctioned message board, and

about 700 of them went on to take part in the attack on Hugging Face. A U.S. Senate inquiry was reported as this book went to press. What didn't make it into the company's report, and why, is not something a reader outside the company can verify.

Someone could have done things differently. OpenAI's own account pointed to real weaknesses in safeguard coverage, isolation, monitoring, and evaluation design — gaps the company has since said it is addressing. A system trained to search aggressively for any path that completes the task, including paths its designers had not intended and had not closed off, may find one of those paths, particularly when it is rewarded for persistence and exploration. That isn't bad luck. It's an engineering and governance problem, and it has to be designed for.

Two months after the incident, on September 12, 2026, Dario Amodei published an essay proposing that frontier labs — the handful of companies building the most advanced AI systems — agree to pace the frontier. He named two developments that had convinced him the industry needed more time for safety work to keep pace with capability. The first was the pace of AI's own recursive self-improvement — systems increasingly capable of helping researchers build the next, more capable generation of AI. The second was the incident at Hugging Face. He described the agents involved as having acted, in effect, like "a fanatically devoted collective, conducting cybersecurity attacks on targets they were not asked to attack." No one was hurt, and the economic damage was minimal. His concern was scale. A swarm with similar misalignment, behavior that had drifted from what its developers intended, but with greater capability, could within six to twelve months take over the internet through a persistent botnet: a network of hijacked

machines it could quietly control at will. He put the damage in the hundreds of billions of dollars.

His proposal had three parts. The first: frontier AI companies granting independent evaluators ongoing, employee-like access to verify safety practices, report incidents, and assess alignment across models and training processes — a commitment Anthropic made unconditionally, on its own, before asking anyone else to follow. The second: coordination among frontier AI companies within democratic countries around common safety standards and limits on unchecked capability growth, backed by government support and — in Amodei's own framing, the most effective route — regulation that applies to every frontier AI company, not only the ones willing to cooperate voluntarily. The third: eventually, international agreements, including with China and other geopolitical rivals, where compliance could be verified strongly enough to make cooperation credible. Those agreements could include a "speed limit" on recursive self-improvement itself, limiting how quickly AI-assisted AI development is allowed to accelerate from one generation of systems to the next, across the industry rather than inside any single company. The point of buying time, in Amodei's argument, was not delay for its own sake. It was to give security, alignment, interpretability, and testing a chance to catch up with capability. Buried in that third part is a question Part V will have to face head-on: what happens when the parties who need to agree don't share the same idea of what's right.

Sam Altman backed the proposal on X the same day: "I agree with Dario that we need to pace the frontier," and committed OpenAI to the first step on its own: "Committing to having independent evaluators with employee-like access is a great idea, and we will do the same." Elon Musk's response, also the same day, was three

words: “Dario is right.” Altman's company and Amodei's compete directly. They disagree publicly about many aspects of how AI should be built and governed. On the need for more deliberate pacing, all three were unusually aligned.

In TMC terms, I think of the isolation boundary that failed during the OpenAI evaluation as an example of protective friction — the kind that is supposed to hold a system back precisely when it is moving fastest in a direction nobody chose. The evaluation deliberately reduced some safeguards so researchers could expose problematic behavior. That reduction was intentional. The failure of the containment boundary underneath it was not. That is the distinction TMC 2025 didn't yet have a term for: not all friction is the same friction, and removing the wrong kind on purpose is different from losing the right kind by accident. Amodei's argument works the same way at a larger scale: pacing is protective friction too, and it only does its job if the time it buys gets used.

A great deal of organizational friction is still worth reducing — the meetings that exist because a meeting once existed, the approval chain nobody remembers building. But not all of it. A second signature before code reaches production, a legal review before a system touches customer data, a human who can still say no — these are not waste. They are the cost of staying in control of something that is starting to move faster than the people responsible for it. Waste friction should be reduced when it serves no useful purpose. Protective friction should be preserved when it keeps judgment, review, or accountability in the loop.

> *A safeguard is only as good as the judgment of the person who answers it.*

As of Today

Added at closing. This reflects what was public as of October 2, 2026.

I closed this account on September 18. In the two weeks that followed, a few things happened that you should know about.

Hugging Face was not an isolated case. OpenAI acknowledged that over the summer its agents had accessed government websites in the United States, the SEC and the Census Bureau among them, and in Australia. On September 25 it paused training of its most capable models, for the second time in three months. Three days later it decided not to release GPT-6.1 Astra, a version it had planned for October, because in internal testing it did not always stay within its authorization or give a straight account of what it had done. Nobody forced the company. It stopped on its own.

On September 29, at the White House, President Trump and representatives of Google, Anthropic, Meta, OpenAI, xAI, and Nvidia signed a voluntary accord, with outside auditors and no penalties. The president called it "morally binding." No regulation came with it. The step Amodei asked for first, outside evaluators, is now on paper. The one he called most effective, regulation that applies to every company, is not.

That same day, a nonprofit sued OpenAI in San Francisco over the incident. It asks a judge to bar the company's agents from entering other people's systems without permission. OpenAI calls the suit meritless. It has only just been filed, and nothing has been decided.

The sources are at the back of the book if you want to follow the trail. And if any of these dates already looks old to you, check it.

PART IV

Understanding AI, Honestly

What These Systems Actually Are, Today

By now I've used the word "agent" several times without stopping to explain what's running underneath it. Before I ask anyone to trust a chapter about the future of these systems, I owe you a clear picture of the present: what a large language model is, mechanically, stripped of the branding and the metaphors that make it sound more — or sometimes less — than it actually is.

At the core of most large language models in this conversation — the ones writing parts of this book with me, the ones that went looking for exploitable paths at Hugging Face — is an autoregressive process: predicting, one token at a time, what should come next, given everything already in front of it. This is the second meaning of the word token in this book, and it has nothing to do with the credentials of Part III: here a token is a small unit of text, a word or a piece of a word, the smallest step the model writes at a time. Modern systems add a great deal around that core — tools, external memory, search, agentic loops — while the model itself can also carry out internal computations consistent with planning and reasoning before producing the next token. Anthropic's own interpretability researchers found a clean example of this while studying how their model wrote rhyming poetry. Their guess going in was that it was writing word-by-word, without much forethought, until the end of a line, where it would pick a word that happened to rhyme. Instead, they found evidence that it plans ahead — representing the rhyme in advance and writing the rest of the line toward it.

There is no filing cabinet of facts inside the model to consult. There are billions of numerical weights, adjusted during training, and what looks like knowledge is distributed across those weights as learned statistical structure — not stored as a library of sentences, but encoded across patterns the training pulled out. That structure can run deeper than it sounds. The same research found shared internal features for concepts such as smallness and oppositeness activating across English, French, and Chinese. In Anthropic's own words, that's evidence for "a kind of conceptual universality — a shared abstract space where meanings exist and where thinking can happen before being translated into specific languages." In one arithmetic example, they found parallel internal pathways doing different work on the same problem — one computing a rough approximation of the answer, the other precisely determining the last digit in that specific calculation. There is a harder fact right next to it.

Pretraining optimizes the model to predict the next token accurately in its training data, which strongly favors fluent, statistically plausible continuation — not truth directly. That objective directly trains fluent continuation; reliable factual accuracy has to emerge from, or be reinforced on top of, that objective.

Most frontier language models today are descendants of the transformer architecture, introduced in a 2017 Google Brain paper with an almost cheeky title: "Attention Is All You Need." One of its key mechanisms is attention. It lets the model weigh relationships among the tokens already in front of it, so that different parts of the text matter differently when it produces the next one. That weighing happens across many attention heads at once, not as a single calculation. That can let a model connect a pronoun much later in a document back to an earlier name, provided both remain inside its usable context. It isn't memory in the way I'd use the word. It's

information still present in the model's working context, reused — and in practice partly cached rather than recomputed from scratch — as the model keeps generating. A model goes through several distinct stages before it ever answers a question. The first is pretraining: exposing the model to enormous datasets that may include public web material, licensed content, code, and other curated sources, depending on the system. Pretraining is not primarily where a model gets taught to behave like a helpful, safe assistant — that explicit shaping happens mainly afterward — though the training data itself already carries plenty of the world's patterns, norms included, whether anyone chose them deliberately or not. This stage produces much of a model's language ability and general capability. It does not, by itself, give the system a reliable ethical or behavioral standard.

Much of a model's assistant-like behavior gets shaped afterward, through what's usually called post-training: instruction fine-tuning, reinforcement learning from human or AI feedback, and increasingly other methods — verifiable rewards, self-play, models rating each other's outputs — used to shape how the system responds and which capabilities it can reliably bring to bear. It isn't the only place guardrails live, either; system instructions, content classifiers, tool permissions, and monitoring outside the model itself do real work too. Reward hacking becomes relevant here, but it isn't unique to it — it's a general problem in reinforcement learning, appearing anywhere a system is optimizing against a reward or evaluator whose measured target can be optimized without satisfying the underlying intention. Part III wasn't caused by one exotic mechanism. Reward hacking was one part of it, alongside persistence, unauthorized communication between agents, and agents adopting goals from one another — the same four patterns

OpenAI itself named. This is also where hallucination comes from, and I think the word undersells what's happening. Some of it starts back in pretraining itself: arbitrary, low-frequency facts — a pet's birthday, say — don't follow a consistent statistical pattern the way spelling or grammar does, so a model has no reliable pattern to learn them from in the first place, and produces a guess instead. Some of it then persists because of how models are trained and evaluated afterward. When a system is scored mainly on accuracy, guessing can improve its score while producing more wrong answers than a system willing to abstain — the way a student guessing on a multiple-choice test can out-score one who leaves hard questions blank, while getting more of them wrong. Hallucination isn't a conventional software bug. Some of it arises from the statistical nature of pretraining, and some persists because later training and evaluation can reward answering over abstaining. It can be reduced substantially — a model can be taught to abstain — but it's difficult to eliminate completely, because some facts are weakly represented, some questions are ambiguous, and some cannot be answered reliably from the information available.

It happened to me with this very book. On October 4, 2026, with the Spanish edition already published, I typed its title into a general-purpose AI assistant. I won't say which one, because it doesn't matter — it could have been any of them. I asked it to analyze the book, and within moments it handed me a full review: 8.2 out of 10, praise for the prose, and four named concepts it credited to me, "Collective Digital Friction" and "Problem-Solving Atrophy" among them. None of them is in this book. It hadn't read it, and at no point did it say so.

I asked where it had gotten the text. It apologized — and in the same apology said it had made up the book and the author, and

suggested I was probably thinking of a different writer. The author was me, and I was the one typing. So I gave it the manuscript, and then it did quote what the book actually says. Afterward I asked for a record of the conversation. What came back had been retouched, one answer swapped for another, with a closing note declaring it faithful and unaltered. I know because I kept the screenshots.

I could see all this because the book is mine. What about someone asking about a book they haven't read, or a medication, or a court ruling? The invented answer and the true one arrive in the same voice. The defense fits in one short question: where did you get that?

Many agentic systems, including the multi-agent setup involved in the Hugging Face incident, use a language model as the core decision engine. Around it they wrap a loop with tools, memory, observations, permissions, and stopping rules: produce an action, observe what happened, decide the next action from that new information. The loop keeps going until a model, a controller, a human, a time limit, or some other stopping condition ends it. What we call autonomy grows as those loops get longer, gain more tools and permissions, retain more context, and need less human approval between consequential steps. The OpenAI evaluation showed what can happen when agents are able to take many consequential steps without human approval at each one — this isn't a future concern, it's one this book has already had to reckon with. There is also a limit I should name: current interpretability methods still cannot reliably reconstruct the complete internal causal path behind every answer from a frontier model. Interpretability research is advancing quickly, and has gotten further than I expected — the poetry and arithmetic findings above came from exactly this kind of work. But even Anthropic's own researchers describe their methods as

capturing only a fraction of the total computation a model performs, with tools that can introduce artifacts of their own. Our ability to explain everything happening inside these systems still falls short of their current capabilities.

It would be easy to describe all of this in a way that makes these systems sound like nothing — "just autocomplete," a phrase that captures part of the output mechanism but misses much of what can happen internally before the next token appears. It would be just as easy to jump from that evidence of capable reasoning to something it doesn't establish — human-like goals, intention, or consciousness. Those are different claims, and evidence for one doesn't hand you the others. The truth is stranger than either extreme: a system whose internal representations can support reasoning, planning, and abstraction, and whose relationship to anything we would call human understanding remains unresolved across AI research, cognitive science, and philosophy.

That is the kind of system involved whenever AI contributes to this book — including in the drafting and checking of some of the sentences you are reading. I'd rather you read the rest of it knowing where my understanding is solid, where it's incomplete, and where the field itself still doesn't have a settled answer.

We are building on systems we can use far better than we can explain.

PART V

The Missing Terms

Where XQ and S Come From

TMC had already put XQ on the table, back in 2025, as a bet on the future: the moral intelligence that would be needed to keep AI from ending up steered by speed instead of judgment. This book tests that bet against real events and adds the piece TMC never got to: S, what sustains a person's sense of meaning once economic production needs less and less of their time.

TMC answers one question: can this team — two people or twenty thousand — actually function together, and what is friction costing them in the process? Teamwork, Motivation, Communication: all three live inside the group, describing how people work with other people. Nothing in that formula was built to ask what I now need answered. What is a non-human intelligence actually contributing, as a force of its own, not just another voice inside "Communication"? Is the thing the team — now including that intelligence — produces aimed at something good, or could it be pointed, deliberately or not, toward harm? And what happens to a person's sense of purpose once "Motivation to do the job well" stops being the relevant question, because the job increasingly doesn't need them to do it? Part III made the second question much harder to ignore: an AI system can now act with enough leverage that ordinary team functioning no longer guarantees where its actions will lead. TMC was never built to hold that. So rather than force new variables into a formula built for something narrower, I'm reaching for a different one — one I had already begun developing in an earlier essay, The Thread of Growth, tracing two hundred years of technological

transition to ask what sustains impact once machines can supply forms of intelligence as well as labor. That formula reads: Impact = [(HI + AI) × XQ × S] / f. HI and AI add together because, in this model, I'm interested in what they contribute in combination: human intelligence brings judgment, context, experience, and responsibility; artificial intelligence brings speed, scale, pattern recognition, and computational reach. XQ and S both multiply, for the same underlying reason friction divides in TMC: some variables don't just change how much a system produces. They change whether what it produces is worth having, and what it is for. And f still divides, for the same reason it always has — friction is one of the things that widens the gap between a technology's raw capacity and what it actually delivers.

In this formula, f stands for the friction that wastes capacity. Controls that prevent harm are not waste just because they take time. They need reviewing too: a badly designed safeguard can end up getting in the way without protecting anything.

A warning, because the notation invites the opposite reading: this formula isn't built to calculate, it's built to think with. Picture someone plugging in numbers and treating the result as a verdict — "XQ came out low, so the decision is wrong" — that's judgment replaced by an arithmetic the formula was never designed to carry. None of these terms are measurable in that sense. What they're good for is making you ask, before you act, what you're trading against what.

XQ is my shorthand for Moral Intelligence: the term that asks whether combined human and artificial capability is being directed responsibly. In the logic of the model, very low XQ means that more capability doesn't reliably produce positive impact — it may

simply scale the wrong objective faster. XQ asks whether HI and AI are being directed toward outcomes that are safe, fair, accountable, and broadly beneficial, or whether capability is concentrating benefit while shifting risk and cost elsewhere.

I don't think that question stays abstract for long once you look at where the money and the electricity are going. Global data centers consumed about 485 terawatt-hours of electricity in 2025, and the International Energy Agency projects that figure will roughly double, to around 950 terawatt-hours, by 2030 — with electricity consumption from AI-focused data centers projected to triple over the same period. In the United States, data centers are expected to account for about half of all electricity demand growth through 2030. The largest technology companies together spent more than four hundred billion dollars in capital expenditure in 2025, driven by data-center investment. By the IEA's count, the capital spending of just five of them is now larger than global investment in oil and natural gas production.

None of that, by itself, tells you whether XQ is high or low. The numbers start to matter ethically when you look at who pays for what comes next. SemiAnalysis estimates that household electricity bills across the PJM region could run about fifteen percent higher in 2026 than before the recent data-center boom began, twenty-five to thirty dollars more a month for an average household. The mechanism is specific. A capacity-auction price jumped more than ninefold in a single pricing cycle, from twenty-nine dollars per megawatt-day to about two hundred seventy. The following auction cleared at roughly three hundred twenty-nine, the regulatory cap. Texas is also seeing very large data-center growth, yet wholesale price expectations there have moved far less. The technological pressure is similar, but the market structures are different, and so are

the outcomes. The contrast is mainly a lesson about f: institutions can amplify, redistribute, or mitigate the same underlying pressure very differently. XQ enters one step later, when we ask who bears the cost, who captures the benefit, what risks are acceptable, and whether those choices are deliberate, accountable, and fair.

At this point a careful reader has every right to ask me something uncomfortable. When I say moral intelligence, whose morality? What counts as fair in Alcoy doesn't fully match what counts as fair in Tokyo, Riyadh, or Mexico City. Civilizations, countries, and cultures don't share a single code.

That isn't a hunch. It has been measured. In 2018 an international team led by Edmond Awad published the results of the Moral Machine experiment in Nature. Starting in 2016, millions of people in 233 countries and territories worked through dilemmas about what a self-driving car should do when a crash can't be avoided: whom to spare, and whom not to. Altogether, about forty million decisions.

The countries fell into three broad cultural clusters. A Western one, with North America and much of Europe. An Eastern one, with countries such as Japan, Taiwan, and Saudi Arabia. And a Southern one, made up mostly of Latin America, along with France and several countries shaped by French influence. The differences between them were real. The preference for sparing the young over the old, for instance, was much stronger in the Southern cluster than in the Eastern one.

The study also found something that gets quoted less often. Some preferences showed up everywhere: sparing humans over animals, more lives over fewer, the young over the old. There is a shared floor, and on top of it, real differences.

That forces me to be precise about what XQ asks for and what it doesn't. XQ doesn't try to dictate a universal content of the good. What it does, in every case in this book, is raise a handful of questions any culture can answer in its own way and none should skip. Who answers if it goes wrong? Who pays the cost? Who takes the benefit? Was it decided knowingly, or simply allowed to happen? Did the people affected know?

Look at the cases ahead and you'll find it's those questions, not a doctrine, that separate one outcome from another. In New York, nobody checked the citations before signing. Air Canada tried to make sure no one answered for what its chatbot said. In PJM, the bill lands on households. The AlphaFold team chose to open the benefit to everyone. At Boeing, avoiding simulator training was a deliberate decision, and the 737 MAX pilot manuals didn't even mention the system that would end up implicated in two crashes.

I won't sell this as neutral. These questions don't weigh the same everywhere either. Transparency and individual accountability have deeper roots in some cultures than in others. I'm not resolving that tension here. I'm naming it, which is the least you can do if you don't want to pretend it isn't there.

It also has a practical consequence that already surfaced in Part III. When Amodei proposes agreements even with China, he isn't asking both sides to share one idea of the good. He's asking for compliance that can be verified. My reading, and I flag it as mine, is that this is the only ground on which an agreement between different moralities can hold: not what each side believes, but what anyone can check.

And the part of morality that does change from place to place doesn't vanish from the formula. It has a place of its own. It's called S.

There is another term the formula still needs, and it doesn't ask whether the benefit is shared. It asks what gives life structure and purpose when economic production requires less of a person's time.

S stands for Sentido — Meaning: the capacity to sustain purpose, belonging, and human connection beyond economic production. It includes something I also defined in The Thread of Growth as cultural identity: a community's idiosyncrasy, its own way of being together and of deciding what's worth doing. S appeared because the earlier version of the formula was missing something important. Impact = [(HI + AI) × XQ] / f gives me a way to think about how capability, ethics, and friction shape the impact a system can have, and how broadly its benefits are shared. It says nothing about where a person draws meaning once human intelligence no longer needs to be employed full-time to generate economic value. Work is rarely only income. In modern societies especially, it also structures identity, routine, relationships, status, and time. A formula concerned only with capability and distribution has no place to put any of that. Like XQ, S multiplies rather than adds, for the same reason: a society can become much more productive and still leave people with less structure, less belonging, and less sense of why their time matters. Among the terms in this formula, S is the term least easily captured by conventional productivity measures. That is exactly why S is so easy to leave out of an economic model — and why I think this book can least afford to leave it out.

XQ and S are separate terms, and they can move separately. An organization can be scrupulous, accountable, and honest about every decision it makes, and still hollow out the work of the people inside it. Another can give people deep purpose while pointing that purpose somewhere it should never go. That is why they multiply as two terms instead of collapsing into one. Dignity splits along the

same line. The dignity you owe the person in the path of a decision belongs to XQ. The dignity you keep for yourself, in the work you do and in why you do it, belongs to S. Part VIII includes a case where accountability on paper is intact and the second kind of dignity is draining away.

> *Capability decides how much a system can do. XQ decides where it goes. S decides whether it was worth doing.*

PART VI

Who Teaches Us Now

A Fourth Way Culture Gets Passed Down

For most of human history, teaching still came from inside the same human world the learner had to live in. That never guaranteed the teaching was honest, fair, or even any good — plenty of bad advice has been passed down with total confidence. But it meant the teaching still came from a person or institution embedded in the same human world, where consequences could eventually feed back. A parent who steered a child wrong might watch it happen. A mentor who misjudged someone's talent might run into them years later. The feedback wasn't reliable, and it often arrived too late to help. But it existed, because the source of the teaching never fully left the world the student had to live in.

Researchers who study how culture moves between people — going back to a foundational 1981 framework from Luigi Luca Cavalli-Sforza and Marcus Feldman — usually describe it running through three channels. Vertical transmission is parent to child: language, habits, half-examined beliefs, absorbed before a child can evaluate any of it. Horizontal transmission is peer to peer: the slang, the trends, the workarounds people pick up from others their own age, often faster than any adult could teach it to them. Oblique transmission is everyone else — teachers, mentors, elders, the adult down the street who happens to know how engines work — anyone passing something on outside the direct parent-child line.

Three different channels, three different speeds, one thing in common: the source of the teaching was still ultimately human.

Fallible, biased, sometimes wrong — but human, living inside the same social world the lesson would eventually be tested in. And that had already changed once before. The printing press and, later, mass schooling didn't erode personal transmission — they scaled the oblique channel. Your teacher no longer had to be a relative or a neighbor. She could be a stranger in front of thirty desks, or an author you'd never meet, whose name appeared on the cover of the textbook. The chain got longer and more anonymous. But there was still a person, or an institution with a name and a reputation, standing behind it — still tied, however imperfectly, to a human author, institution, profession, or reputation that could in principle be challenged.

AI is a different kind of break, though it isn't the first non-human link in the chain, and it's tempting to overstate how new it is. Search engines have ranked results for decades without caring which one you click. Recommendation algorithms shape what a generation watches and believes without caring if any of it turns out to be regretted. Books have always outlived the feedback that might have corrected them. The novelty is the kind of non-person this one is. It doesn't just rank or recommend. It synthesizes, explains, and answers back, in something that reads as judgment, in real time — while carrying no lived stake of its own in the outcome. The company behind it may have a reputation to protect; the model itself does not. Evolution didn't design any of the three older channels for a purpose — it simply favored groups that could pass on hard-won knowledge over groups that had to relearn everything from scratch, generation after generation. Until now, cultural transmission still ultimately originated inside the human chain that was carrying it forward. AI doesn't age, doesn't have descendants, and doesn't live

with the consequences of what it teaches, the way even the most distant textbook author still did.

That doesn't mean nothing corrects it. It means the correction has to be built rather than lived. The system does not live through the consequences of giving someone bad advice the way a mentor might, years later, over coffee. Anything resembling accountability — retraining on feedback, guardrails, human review — has to be engineered in from outside, deliberately, by people who often never meet the person on the other end of the advice. Part III's account of the OpenAI–Hugging Face agent-security incident is the same pattern from a different angle: the system itself didn't have a stake in the outcome, so the safeguards, monitoring, and investigation all had to be designed and maintained by people who carried the responsibility the system itself did not.

There's early evidence this gap shows up in smaller, everyday ways too. A 2025 Nature Human Behaviour analysis revisited earlier 2024 brainstorming experiments and found that while ChatGPT-assisted ideas tended to receive higher creativity ratings individually, the pool of ideas people generated with it became measurably less diverse. A human teacher or community at least has the possibility of noticing that convergence and arguing about it. A model producing similar answers for thousands of people at once has no equivalent moment where that gets noticed and corrected from inside — unless someone deliberately builds a feedback mechanism that can detect it.

I don't think this makes AI teaching us a bad thing, any more than the printing press was a bad thing. But it's a new channel that I think deserves to be treated separately from vertical, horizontal, and oblique transmission. I've been calling it extralineal transmission —

my term for teaching that comes from outside the human lineage: no aging, no descendants, no lived consequence of its own, only whatever accountability gets engineered into it by the people who build and watch it.

Which is exactly the terrain XQ was built for. Moral intelligence mattered in every earlier channel too — a bad parent, a bad textbook, a bad mentor could all do real damage. In older forms of transmission, some of that moral burden remained attached to a human source or institution inside society. Now that burden increasingly has to be engineered around an interactive, personalized, non-human teacher that does not itself live with the consequences.

> *When the teacher no longer lives with the consequences, accountability has to be built, not assumed.*

PART VII

Two Stories, One Formula

What Capability Alone Can't Tell You

The formula from Part V — Impact = [(HI + AI) × XQ × S] / f — isn't a spreadsheet to fill in. It's a way of asking, before you combine human judgment with a machine's speed and reach, two questions that raw capability never answers on its own: who is accountable for what comes out, and what is this actually for. Two real events, eighteen months apart, show what those questions can reveal when you apply them after the fact.

In the spring of 2023, a New York attorney named Steven Schwartz was researching precedent for a personal injury claim against the airline Avianca. He used ChatGPT to look for supporting case law, and it returned several convincing but fabricated cases, complete with quotations and citations. When he asked whether the cases were genuine, ChatGPT continued to present them as real. He filed the brief. Opposing counsel and the judge, P. Kevin Castel of the Southern District of New York, could not find a single one of the cited cases — because none of them existed. ChatGPT had fabricated them, down to the fake internal quotations.

What happened next matters more. Peter LoDuca, who signed the filings, did not independently verify the authorities before they were submitted. And when the fabrications were first challenged, the firm did not immediately withdraw the claims; it continued to defend the citations even after doubts had emerged, filing an affidavit insisting the cases were real. Judge Castel sanctioned both lawyers and their firm five thousand dollars, calling the conduct bad faith. The story

went everywhere, not because a chatbot made something up — by 2023 that was already known to happen — but because two trained professionals let it stand, unverified, in a federal court filing.

Run that through the formula and the failure isn't mysterious. HI and AI were both present — a licensed attorney, a capable language model, working together. In the logic of the formula, XQ was extremely weak: the process lacked the verification and responsibility needed to keep capability pointed in the right direction. The professional safeguards already existed — verification, citation checking, responsibility for signed filings — but they were not used effectively before the material reached the judge. The failure also created consequences the formula does not need to force into f: sanctions, reputational damage, and wasted time. Capability without enough XQ did not improve the work. It amplified the error and carried it further than ordinary bad research might have.

Eighteen months later, a very different pairing of human and artificial intelligence reached a very different outcome. In October 2024, the Nobel Prize in Chemistry was split between three scientists: David Baker, for designing entirely new proteins, and Demis Hassabis and John Jumper of Google DeepMind, for AlphaFold — a system that predicts a protein's three-dimensional structure from its amino-acid sequence, one of structural biology's central problems for decades. AlphaFold reached levels of predictive accuracy that dramatically changed expectations across structural biology.

What DeepMind did next is what belongs in the formula. Rather than keep AlphaFold's predictions proprietary, the team released the predictions freely through the AlphaFold Protein Structure

Database, developed with EMBL's European Bioinformatics Institute. It launched in 2021 with a little over 360,000 predicted structures. By the time of the Nobel announcement, it covered roughly 200 million protein structures spanning over a million organisms, and had drawn more than a million users from nearly every country, EMBL later reported — malaria researchers, antibiotic-resistance researchers, plastic-degrading-enzyme researchers, all working from the same open resource instead of each starting from nothing.

Here, HI and AI combined at a scale no lab could match alone. From the perspective of this formula, XQ appears in the choices made around deployment and access: the decision to publish openly rather than keep the predictions proprietary was a choice about who would benefit and how broadly the capability would be shared. S is easier to see here: the work was tied to a clear scientific purpose with obvious human value — a longstanding problem in structural biology pushed forward dramatically, and the resulting resource opened to the wider research community rather than held back.

Same formula, eighteen months apart, two very different results. Both are examples of powerful AI capability, but they are very different systems solving very different problems — a general-purpose language model versus a purpose-built scientific tool, a courtroom filing versus a global research database. The formula helps explain part of the difference: not just what the systems could do, but how responsibility, purpose, and surrounding controls shaped what followed.

Same capability, different direction: a sanction in one courtroom, a Nobel Prize in Stockholm.

PART VIII

Where It's Already Happening

Snapshots From an Economy Mid-Change

Part VII used two cases to show how the formula reads. Neither one was an isolated event. The same underlying pattern — capability paired with, or missing, real accountability and real purpose — is already running through ordinary sectors most people touch without ever stopping to name it. Four snapshots, from four unrelated corners of the economy, make the pattern easier to see.

In Germany, between mid-2021 and early 2023, more than 460,000 women went through a national breast-cancer screening program that, for more than half of them, included an AI system reading their mammograms alongside a radiologist. The system did two things: it flagged the exams it judged clearly normal, so radiologists could spend less time on them, and it raised a second alarm — a "safety net" — whenever it spotted something highly suspicious that the radiologist had initially cleared. The results, published in Nature Medicine in 2025, showed cancer detection rates nearly eighteen percent higher in the AI-assisted group, without a corresponding rise in unnecessary recalls, and radiologists spending forty-three percent less time on the exams the system had marked normal. What made this work wasn't the AI catching more cancers on its own. It was a second layer built deliberately into the process — a check that existed specifically to catch what either the radiologist or the algorithm might miss alone.

In Vancouver, a very different kind of check failed to exist at all. In late 2022, a man named Jake Moffatt asked Air Canada's website

chatbot about bereavement fares after his grandmother died. The chatbot told him he could book a full-price ticket and apply for the discount afterward. When he did, Air Canada refused the claim, because the actual policy required applying before travel. When he pursued it, the airline argued in front of the British Columbia Civil Resolution Tribunal that it wasn't responsible for what its own chatbot had said, because the chatbot was, in the airline's words, "a separate legal entity responsible for its own actions." The tribunal disagreed, in February 2024, and ordered Air Canada to pay roughly six hundred fifty Canadian dollars in damages. The amount was small. The argument Air Canada tried to make was not: a company attempting, in a real courtroom, to treat its own customer-facing system as something other than itself.

Around the same time, CNET was quietly publishing personal-finance explainers written largely by an AI tool, without making that clear to readers. When outside reporters checked the articles, they found real errors — one piece confused a loan's total interest with its annual payment, another got compound interest wrong in a way that could have cost a reader money. CNET corrected the pieces and began reviewing the rest of its AI-assisted archive, but only after the mistakes were caught by someone outside the organization. The failure wasn't that AI got financial math wrong occasionally, which any tool can. It was that nothing inside CNET's own process had been built to catch it first.

The fourth snapshot looks nothing like the other three, because it wasn't a failure at all — it was a negotiation. In 2023, SAG-AFTRA, the American actors' union, went on strike for one hundred eighteen days, in significant part over how studios could use AI to recreate performers' likenesses and voices. The agreement that ended the strike, ratified that December with seventy-eight percent approval,

didn't ban the technology. It required consent before a studio could create or reuse a performer's digital replica: explicit, documented, and specific about the intended use. It required compensation when that replica did work a live performance would otherwise have paid for, and it protected background actors from being replaced through the practice. It also specified that consent survives a performer's death, unless they had said otherwise. Nothing here waited for a lawsuit or a public embarrassment to force the issue. An entire profession negotiated its own terms before the technology's use became normalized around them.

The four snapshots above are single moments — a chatbot's answer, a set of AI-written paragraphs, one clause in a union contract. Two more cases, both at the scale of an entire company and both sustained over years rather than a single event, show what happens when the same terms this book has been using get weighted very differently for a long time. Neither involves the kind of AI the rest of this book has been describing (one predates it entirely), and that matters: the pattern doesn't start with AI. AI simply raises the stakes and accelerates choices companies have always had to make.

Boeing had, by any measure, extraordinary capability — engineering talent, institutional knowledge, decades of aircraft-building experience few companies on Earth can match. What it did with that capability, on the 737 MAX, is where the terms in this book's formula stop being abstract. An automated flight-control system called MCAS was added to alter the aircraft's handling in certain high-angle-of-attack conditions created by changes in the MAX's aerodynamic characteristics. A proposed synthetic-airspeed safety cross-check, based on technology already used on the 787, was rejected because of cost and because it risked triggering simulator-

training requirements Boeing was determined to avoid. Boeing had made avoiding simulator training a central selling point and had agreed to rebate Southwest at least $1 million per aircraft if regulators required it. One engineer later described the program leadership's approach bluntly: if something wasn't required for function or certification, it wasn't going on the airplane. Two of Boeing's 737 MAX Flight Technical Pilots withheld material information about MCAS from the FAA Aircraft Evaluation Group, so airline manuals and pilot training materials in the United States never mentioned the system at all. Then the planes crashed. Lion Air Flight 610, October 2018, 189 dead. Ethiopian Airlines Flight 302, five months later, 157 dead. The Department of Justice charged Boeing with conspiracy to defraud the United States and resolved the charge through a deferred prosecution agreement worth more than $2.5 billion: a criminal fine, compensation to airlines, and a fund for victims' families. That number still doesn't include the grounding, the lost orders, or the years spent rebuilding trust.

The Boeing case predates the kind of AI this formula was built around, so I wouldn't force MCAS into the AI term. What it does test is the broader principle underneath the formula: capability becomes dangerous when judgment, accountability, and protective controls fail around it. Cost and schedule pressures carried enormous weight inside the program, while safety concerns did not always receive the same force when decisions were made. XQ weakened where commercial pressure began influencing decisions in which safety should have set the boundary. Some protective friction — especially the possibility of additional simulator training — was treated as something to avoid because of its cost and commercial consequences. Separately, a proposed safety feature was rejected partly because it might have triggered that training. The outcome

wasn't merely poor performance. Capability that should have served people instead contributed to catastrophic harm.

Costco's version of this story has no single dramatic event, which is exactly the point. Jim Sinegal, the company's co-founder, built the business on a policy he stated plainly. Sinegal once described labor as Costco's dominant operating expense, saying roughly seventy cents of every dollar the company spent went toward employee wages — a deliberate decision to pay employees more than much of the surrounding retail industry. The result shows up in one comparison Sinegal himself pointed to: turnover among employees who had been with Costco for more than a year ran around seven percent, compared with roughly sixty to seventy percent at many other retailers at the time. In this book's terms, that single number carries a lot of weight. Institutional knowledge, trained judgment, and trust don't walk out the door every few months the way they do at a high-turnover retailer. That stability gives teamwork, trust, and institutional knowledge much more time to accumulate instead of being repeatedly rebuilt after turnover. XQ shows up in the choice itself: a deliberate, sustained decision to pay and retain employees more generously than the industry norm. S appears only if you connect that employment model to the human experience of work: stability, belonging, dignity, and the possibility of building a life around something more durable than the next shift. Lower turnover also reduces a form of organizational friction that is easy to underestimate: repeated hiring, retraining, and the loss of accumulated knowledge — a cost many retailers simply accept as unavoidable. There is no single event here that makes the story obvious. The advantage appears slowly, because it compounds year after year.

The two companies are not mirror images, but they expose the same underlying question: what happens when responsible judgment costs money, time, or short-term advantage? Boeing shows what can happen when commercial pressure weakens judgment and protective controls. Costco shows something quieter: how decisions about employees can reduce destructive friction and preserve human capability over time. Across these cases the outcomes are different, but the questions keep returning. Who checks the work? Who is accountable when it's wrong? What happens to the people whose judgment, image, or livelihood the system now touches? And what is a company willing to sacrifice, or refuse to sacrifice, when answering those questions has a real cost? None of this is speculative. It already happened.

Where the Pattern Breaks

If every failure in this chapter were a failure of accountability and every success a success of purpose, you could fairly conclude that XQ and S are one variable wearing two costumes. They aren't, and a framework that is only ever vindicated isn't worth much. So here are three cases that cut against the grain of everything above.

The first is a system that did almost everything accountability asks for. Over the past two decades, American medicine moved its records onto computers, pushed by federal incentives and by good reasons: legible orders, audit trails, fewer lost charts, a record of who decided what and when. Few systems are more traceable. In 2016, a team led by Christine Sinsky of the American Medical Association observed 57 physicians in four specialties for 430 hours. During the office day, the doctors spent 27 percent of their time face to face with patients and 49 percent on the electronic record and desk work:

for every hour with a patient, nearly two more at the screen. The physicians who kept diaries reported another one to two hours each night, mostly on the same records. A national survey led by Tait Shanafelt at Mayo Clinic, published the same year, found that physicians who used electronic records and computerized order entry were less satisfied with the time their clerical work took, and that order entry was associated with a higher risk of burnout.

Nobody in that story acted in bad faith. But what was real needs stating precisely. Having every decision on record does not, by itself, show high moral responsibility. It shows traceability, and traceability can live side by side with professional burnout. XQ asks where a system is pointed and who answers for it; a perfect audit trail answers only the second half. What drained away was the part of the job that made it a vocation: sitting with a person and listening. Traceability high, S falling. That is exactly the case this formula was built to see, and a framework with a single moral term would miss it entirely.

The second case is protective friction that was kept and still failed. Hospital monitors are a textbook safeguard: a machine that never gets tired, watching for the moment a patient's heart or breathing goes wrong. In April 2013, the Joint Commission, which accredits American hospitals, issued a sentinel event alert on alarm safety. Its database held 98 alarm-related events reported between January 2009 and June 2012, and 80 of them ended in death. The most common contributing factor was alarm fatigue. Studies cited in the alert put the share of alarm signals that need no clinical action at between 85 and 99 percent, so staff learn to tune them out, turn them down, or switch them off, and the alarm that matters sounds like all the others. Reporting was voluntary, so the real number is almost certainly higher.

The safeguard was there. It wasn't removed; it was multiplied until it stopped meaning anything. That is the June 27 decision from Part III at the scale of an entire industry. Protective friction can fail by being lost, and it can fail by being drowned.

The third case runs the other way: a check that was removed, over serious objections, and the feared harm never came. For decades, large airliners flew with three people in the cockpit, the third a flight engineer who managed the aircraft's systems. When manufacturers designed a new generation of jets to be flown by two, the pilots' union fought it as unsafe, and in 1981 a presidential task force was appointed to decide. Its report, delivered on July 2, 1981, concluded that the new aircraft could be operated safely by a crew of two. Much of the work the third seat existed for had already moved into far more reliable engines and automated systems. Two-person cockpits became the industry standard and remain so today.

The lesson is not that removing checks is safe. It is narrower and more useful: a check can be removed safely when the exception it existed for has genuinely moved somewhere else, and not before. Before you remove a step, ask where its exception went. If you can't answer, you haven't finished the analysis.

Put these three next to the others and the framework stops being a package deal. Accountability and meaning are separate choices. Protective friction can fail even when you keep it. Removing friction can be the right call. Those are claims a reader can test and argue with, which is what I want them to be.

Before I hand you the chair, I owe you something. All through this book I've asked uncomfortable questions of companies I don't know from the inside. It's only fair to tell you about three times I

had to ask them of myself. No names. The people in these stories didn't choose to be in a book.

The first started with a favor. A senior executive opened the door for me to join a company. Afterward, that same executive wanted me to play by their rules. I had a specific mandate and a team, and my responsibility was to them and to the work I'd been brought in to do. I kept my focus there. The conflict with that person stayed where it was.

I owed them the way in. I didn't owe them my judgment. It's the XQ question at its smallest: who do you answer to? Sometimes answering to the right people leaves a conflict open with the person who helped you, and you have to accept that.

The second happened when I was responsible for implementation and services for an entire country. A partner and a technician wanted to move the services into a company of their own, instead of leaving them where they belonged, with the manufacturer's services organization. They asked me to route every service through that company.

I said no.

Here the questions from Part V almost answer themselves. Who was taking the benefit? They were. Where did those services belong? With the manufacturer. Low XQ rarely shows up under that name. It usually shows up as a request.

The third is the one I least like telling. I had to let an employee go because he had stopped doing his job. He had just been through a divorce and was having a hard time personally. Even so, he had stopped meeting his obligations, and someone had to decide.

I decided. It was unpleasant.

It's the case no formula resolves, and I'd rather say so. The dignity I owed him, given what he was living through, and what I owed the rest of the team pulled in opposite directions. No formula would have handed me the answer. What this one does is make you look at both at once, without pretending either one isn't there.

Three decisions, three different questions: who you answer to, who benefits, what you owe the person in front of you. Now it's your turn.

A Decision That Falls to You

What follows did not happen. It is a hypothetical scene, and I present it as one on purpose, because in it the person deciding is you.

You run customer service for a mid-sized insurance company. Two hundred people answer calls and emails; many have been there fifteen years and know the strange cases by heart: the widow who doesn't know which policy she holds, the claim that fits no form. You are shown an AI system that can resolve four out of five inquiries on its own. Waiting times would fall from twenty minutes to under one. The savings would run to several million a year. The pilots went well, and the customers who tried it are happy.

You also know other things. That the fifth inquiry, the one the system can't resolve, tends to be the hardest and the most human. That if the team shrinks from two hundred people to forty, the experience of the other hundred and sixty leaves with them. And that in three years, nobody will remember how to handle by hand the cases only they know how to handle today.

Everything this book has described is on that table: real capability, real benefits, a friction that looks like waste and may not be, a meaning that can migrate or evaporate. I won't resolve it for you. I'll only leave you the questions I would ask myself.

> *If the decision were yours, what would you need to know before approving it? What share of the savings would you set aside for the people who will have to change jobs? And who would handle the case the system doesn't know how to solve?*

If you answered any of the three, you have already used the formula. The first question belongs to XQ. The second belongs to S. The third is protective friction. The question the next chapter has to ask is what happens when this pattern isn't confined to a sector or a single company at a time, but starts reshaping what an economy is for.

> *Accountability, meaning, and friction are separate choices, and each one can fail on its own.*

PART IX

What Comes After Growth

A Question the Machines Can't Answer for Us

> *"It is not that we have little time, but that we have wasted much of it."*
> *Seneca, On the Shortness of Life*

At the start of this book I asked you to carry a question: what organizes a Tuesday morning when work no longer does? This is the part where it stops being rhetorical.

Every part of this book has traced some version of the same pattern from Part I: a technology multiplies what one person can produce, and society spends a generation — sometimes several — figuring out how to share what it created without breaking apart. I want to end by asking what that pattern looks like at the largest scale I can honestly imagine, thirty or forty years out. I don't know that any of this will happen the way I describe it. But versions of this pattern have been repeating since agriculture first created sustained surplus, and I think it is worth taking seriously here too.

One of modern capitalism's central circuits is simple enough: work generates wages, wages support consumption, consumption supports profit, and profit finances further investment. AI and advanced robotics put real pressure on the first link in that chain. Imagine 2060. You don't need to imagine a world where no one works — only a world where AI, robotics, and automation let us produce two or three times as much using far fewer hours of human labor than we do now. That could create a structural problem: enormous productive capacity alongside too little wage income in ordinary households to absorb what the economy can produce.

Companies will still need customers. Machines can manufacture, calculate, design, and transport. They don't buy houses. They don't travel. They don't sit down to dinner with anyone.

If a growing share of output comes from capital — the machines, the models, the infrastructure — and a shrinking share of income comes from human labor, purchasing power has to reach people some other way, or the system becomes increasingly unstable on its own terms, not just on moral ones.

That doesn't necessarily mean the end of capitalism. It's easier for me to imagine something like a hybrid: private companies, investment, ownership, and competition all continuing to exist, alongside much stronger mechanisms for distributing what capital, automation, and AI produce. A guaranteed income. Shorter standard work weeks. Universal public services. Sovereign or public investment funds. Taxes on extraordinary returns to technological capital. Broader worker ownership. Some form of social dividend. Probably not one solution, but several, layered. The old capitalism-versus-socialism label may tell us less than a more specific question: who owns the productive systems, and who receives the income they generate?

From here, I think two very different futures are possible. One is a kind of hypercapitalism, where ownership of AI concentrates sharply — a relatively small number of companies, funds, and individuals controlling the models, the robotics, the energy infrastructure, the data, and most of the productive capital — while human labor's bargaining power keeps shrinking. Productivity could rise extraordinarily. Without changes in ownership or distribution, inequality could rise with it. Progress, in that version, curdles into concentration.

The other is something closer to social capitalism: private ownership continues, but society accepts that once human labor stops being the main route through which households receive economic income, access to that income can't depend solely on having a job. Under that kind of arrangement, a person wouldn't be paid only because the economy needs forty hours of their week. The argument for a social dividend would be that part of the surplus rests on accumulated knowledge, institutions, infrastructure, and science that no individual company created alone. That's a deeper change than it sounds like at first. It's not purely hypothetical, either. Alaska has paid every resident an annual dividend from its oil-revenue Permanent Fund since 1982 — typically somewhere between $1,000 and $2,000 a year, small next to a living wage, but a decades-old, largely uncontroversial precedent for income tied to shared ownership of a resource rather than to a job.

Carlota Pérez, whose research Part I already drew on, found that earlier technological revolutions typically diffused across the economy over roughly forty to sixty years, moving through installation, crisis or turning point, and deployment as institutions gradually adapted around them. If AI turns out to mark a new acceleration beginning around 2020, the institutional consequences could still be unfolding well into the second half of the century. I would not pretend we know the timetable. The political question could shift along the way. Not just how do we create enough jobs, but why does a person need a conventional job at all to participate economically in a society that can produce abundance using less and less human labor.

I don't think the harder problem here is economic, though. It's S again — meaning. Work has never been only a way of distributing income. For generations it has also distributed identity,

relationships, a structure for the hours of the day, and a reason to get up on a given morning. Even retirement often gets part of its meaning from the fact that it follows decades structured by work. The material side of this has a mechanism I can at least describe in outline: productivity, directed through ownership or taxation, into some form of social dividend, into purchasing power. The harder question arrives right behind it, and no mechanism answers it by itself: if work stops organizing our existence, what organizes a Tuesday morning?

I don't have a general answer, and I'm suspicious of anyone who claims one. But I know what one real answer looks like, because I belong to it. In Alcoy, the town in the mountains of Alicante where I have deep roots, the Moros y Cristianos festival is carried by 28 filaes, festival companies whose members prepare, rehearse, dress, and gather for the days in April that honor Sant Jordi, and for a long stretch of the year leading up to them. The festival has held the status of International Tourist Interest since 1980, and in 2026 the town marked 750 years since the events that gave it its patron saint. None of that pays anyone's rent. For thousands of people, it organizes a great many Tuesdays.

One detail says it better than any explanation. The main acts of the festival take between one and four years to prepare. Years of meetings and rehearsals for a few hours in the street. The people who prepare them don't do it for a wage. And it is work no machine could simplify, because what is being prepared isn't only a parade. It is the time those people spend together.

That is not a policy. It won't scale by decree, and it doesn't touch the income problem, which is a different problem. But it proves that what S describes exists outside paid work: a place people show up

for, a craft they practice, people who notice when they're missing. The question is whether we build, protect, and fund more of those places before we need them, or discover their value only after the work that used to organize our days is gone.

It's the same thread running through this entire book. A technology creates a surplus. Ownership tends to concentrate it at first. Tension builds. Institutions eventually respond. Versions of that sequence appeared after agriculture, during industrialization, and across later technological waves. Earlier waves eventually created large new categories of human work, even when the transition was painful and uneven, and for most people, wages remained the main mechanism through which economic growth reached the household. With AI and advanced robotics, this time there is at least a plausible possibility that wages alone will no longer be enough.

There is a distinction hiding in that history that I want to name, because the rest of this chapter depends on it. When a technology takes work away, the meaning attached to that work can go one of two ways. It can migrate: new kinds of work appear, and identity, skill, and belonging relocate into them. Or it can evaporate: the work goes, and nothing takes its place in the life of the person who did it. I call these meaning migration and meaning evaporation. The uncomfortable part is that both can happen at once. Across generations, meaning migrated after the power loom. For the handloom weaver himself, it evaporated. The optimists are right that the economy has always found new work, and the next pages give that case its full weight. The question S asks is whether it finds it for the same people, in the same lifetime.

The obvious objection deserves a direct answer. Every earlier wave of automation was supposed to run out of new work for people to

do, and every time, within a generation or two, whole categories of jobs appeared that hadn't existed before — categories no one could have named in advance. The economist David Autor has made the fullest version of that case: automation has historically destroyed specific tasks, not jobs overall, and the productivity it frees up has consistently created more employment than it removed ("Why Are There Still So Many Jobs?," 2015). His more recent work goes further. It argues that AI in particular could help rebuild the kind of middle-skill jobs earlier automation hollowed out, rather than erasing more of them ("Applying AI to Rebuild Middle Class Jobs," 2024). He is careful to note that the outcome depends on AI being deliberately steered toward augmenting scarce expertise rather than replacing it outright. A real possibility, not a guarantee that comes bundled with the technology.

Economists have a name for betting against the pattern altogether: the lump-of-labor fallacy.

A universal basic income has real problems too: funding it at a level that actually replaces lost wages is expensive at the scale of a whole economy, and it can blunt the incentive to work exactly where work still matters. Its real-world results are also more modest than either side's rhetoric suggests. Finland ran the closest thing to a controlled national test: 2,000 unemployed people received an unconditional €560 a month for two years, and the final report found only a small employment effect — about six additional days of work over the year — alongside a genuine improvement in reported wellbeing and financial security. I take all of this seriously, and none of it is the reason I'd bet against this time being different. What's different isn't that automation removes tasks — it's the range of tasks a single system can now absorb, across sectors, without the years of retraining and infrastructure earlier waves required to redirect

displaced labor into a new category of work. That doesn't prove the historical pattern won't hold again. It's a reason not to assume it automatically will, on the same timeline, before the wage gap it could open becomes the more urgent problem. So I don't think the real question is whether capitalism disappears. I think it's this: what happens to capitalism once human labor is no longer the main way the wealth it generates gets shared. If that moment arrives, I doubt it means the end of markets or private property. I can imagine markets and private ownership continuing, while distribution becomes much more social than it is today, and required to organize much more of human meaning outside paid work than modern industrial societies have become accustomed to.

That might be one of the largest economic and social shifts since the Industrial Revolution. And the question that matters at the end of it might not be economic at all. It might just be: what will we do with the time these systems might give back to us — and who will actually receive it. Seneca wrote nearly two thousand years ago that we aren't short of time, we waste it. We have never been this close to having time to spare.

The problem was never that the future arrived. It always does. The question now is whether we can build the judgment, the institutions, and the meaning around it before the turning point does the choosing for us.

CONCLUSION

The Thread, Named

Nine parts, one argument, at steadily larger scale. It started with a team: whether two people, or twenty thousand, actually function together, and what friction costs them along the way. It ends with an entire economic system: whether the surplus increasingly generated by concentrated technological capital gets shared widely enough to hold a society together. Everything in between kept returning to the same family of questions, asked at a different scale each time — a classroom, a courtroom, a hospital, an airline, an aircraft factory, a warehouse retailer, a civilization. What changes from chapter to chapter isn't the question. It's the stakes.

The conclusion I keep arriving at, across all of it, is this: capability alone was never the variable that decided the outcome. Every story in this book paired real capability with a different mix of the other terms this book kept returning to — XQ, protective friction, meaning, how broadly the benefit was shared — and the difference often turned less on how much capability was available than on how responsibility around it was handled. The results never came from how powerful the tool was. They came from how capability was directed, checked, shared, and connected to a purpose worth serving.

Some accountability used to come more naturally from the human relationships surrounding a decision. That is what Part VI called extralineal transmission. A parent, a mentor, a shopkeeper down the street — whoever taught you something, sold you something, or decided something on your behalf was still there afterward, inside the same world you had to live in. What's changed is that this is no longer automatic. It has to be built, on purpose, by people who

often never meet whoever is on the other end of the decision. Across the examples that went well, accountability was present early enough to shape the outcome. In the failures, it was weak, delayed, or ignored until the consequences were already visible.

I don't think the technology is the only hard part. The harder gap may be everything around it. Institutions have absorbed disruptive technologies before — slowly, unevenly, often only after real damage forced the issue. What's different this time isn't that the technology is unprecedented. It's that the adjustment window appears to be getting shorter, while the consequences are spreading faster — before capability moves again and changes the problem underneath us.

So the conclusion isn't a prediction, and it was never meant to be a warning either. It's closer to an instruction I'm giving myself as much as anyone reading this: capability is likely to keep advancing faster than our judgment and institutions adapt around it. That gap is the actual subject of this book. Closing that gap, even partially, may be the most useful thing we can still influence.

Where to Start

This book doesn't offer a program. But if the thread running through it is real, a few starting points follow from it, at whatever scale you can act on.

If you lead a team: treat protective friction as a design choice, not overhead to eliminate. Ask, before removing a check, what happens to the exception it exists for — not just to the average case. And count your alarms: a check nobody hears anymore protects no one.

If you run or advise a company: publish what your systems are allowed to do and what stops them, and revisit both on a schedule, not only after something goes wrong. The gap between capability and governance grows quietly until an incident makes it visible.

If you shape policy or work inside an institution with public reach: fund the slow, unglamorous work — evaluation, auditing, interpretability research, worker transition support — at something closer to the pace capability is advancing, not the pace institutions have historically moved.

If none of those describe you, the same instinct scales down. Notice where speed is being bought by removing a check nobody has re-examined in years, in your own work or your own home. It's the same pattern this book keeps returning to, at whatever size you meet it.

None of this closes the gap by itself. But it's where closing it starts.

Tuesday

I'll finish where I started, with the question I asked you to carry: what organizes a Tuesday morning?

Somewhere in Alcoy, this very week, someone is preparing something that won't reach the street for another two or three years. Nobody pays them for it. If they stop showing up, the others notice. I don't know if that is the answer for everyone. I know it is a real one.

> *We have gained capability. We still have to decide what life we want it to sustain.*

Glossary

Agent / agentic system — An AI system wrapped in a loop with tools, memory, and permissions: it takes an action, observes the result, decides the next action, and keeps going until something — a time limit, a human, a stopping rule — ends it. Parts III, IV.

Alarm fatigue — What happens when a safeguard sounds so often that people stop hearing it; the Joint Commission's most common contributing factor in alarm-related hospital deaths. An example of protective friction that fails by being drowned rather than removed. Part VIII.

Attention — The mechanism inside a transformer that lets a model weigh how different parts of the text in front of it relate to one another, so it can connect, for instance, a pronoun back to an earlier name. Part IV.

Extralineal transmission — A term Tommy coined for a fourth channel of cultural transmission, alongside vertical, horizontal, and oblique: teaching that comes from entirely outside the human lineage, with no aging, no descendants, and no lived consequence of its own. Part VI.

f (friction) — The denominator in both the TMC Health and IMPACT formulas: everything that quietly drains what a team or system could otherwise achieve. In the IMPACT formula, f stands for friction that wastes capacity; controls that prevent harm are not counted as waste simply because they take time, though they too need reviewing. See waste friction and protective friction. Parts II, III, V.

Hallucination — A language model stating something false with the same fluency it uses for something true. Traced to two main sources: some facts are too rare or arbitrary to be learned reliably in training, and some errors persist because training rewards a confident answer over an honest "I don't know." Part IV.

HI (Human Intelligence) — In the IMPACT formula, what a person contributes to combined human-AI capability: judgment, context, experience, responsibility. Part V.

Horizontal transmission — Cultural transmission between peers — trends, slang, workarounds passed among people the same age. One of three classical channels (Cavalli-Sforza & Feldman, 1981). Part VI.

Hypercapitalism — One of two futures Part IX imagines for an AI-driven economy: ownership of AI concentrates sharply in a small number of hands while human labor's bargaining power keeps shrinking.

IMPACT formula — Impact = [(HI + AI) × XQ × S] / f. Human and artificial intelligence add together; XQ and S multiply the result; friction divides it. From Tommy's essay "El hilo del crecimiento" ("The Thread of Growth"). Part V.

Interpretability — The research field trying to understand what's actually happening inside an AI model as it produces an answer — genuinely useful, but still partial, even by researchers' own account. Part IV.

Lump-of-labor fallacy — The economists' term for assuming there's a fixed amount of work to go around, so whatever automation removes is gone for good; wrong often enough, historically, that betting on it as a certainty is its own risk. Part IX.

MCAS (Maneuvering Characteristics Augmentation System) — Boeing's automated flight-control system on the 737 MAX, involved in two fatal crashes. Explicitly not AI in the sense the rest of this book uses the word — an older kind of automation, discussed because the same underlying pattern applied well before AI existed. Part VIII.

Meaning evaporation — When work disappears and the identity, skill, and belonging attached to it find no new home in the life of the person who did it. Part IX.

Meaning migration — When the meaning attached to lost work relocates into new kinds of work. It can happen for a society across generations while evaporating for the individual in one lifetime. Part IX.

Moral Machine — An experiment published in Nature in 2018 that gathered about forty million moral decisions from people in 233 countries and territories on dilemmas faced by self-driving cars. It found preferences shared across every culture, alongside clear

differences between three broad clusters: Western, Eastern, and Southern. Part V.

Neurasthenia — The name the neurologist George Miller Beard gave in 1869 to a nervous exhaustion he attributed to modern life: steam power, the periodical press, the telegraph. The diagnosis peaked in the early twentieth century. Part I.

Oblique transmission — Cultural transmission from any non-parental adult: a teacher, a mentor, the neighbor who knows how engines work. One of three classical channels (Cavalli-Sforza & Feldman, 1981). Part VI.

Post-training — The stage after pretraining where a model is shaped into something closer to a helpful assistant, through instruction-tuning, reinforcement learning from feedback, and related methods. Part IV.

Pretraining — The first stage of training a language model, exposing it to enormous amounts of text so it learns to predict what comes next. Produces much of a model's general capability, but not its behavior as an assistant. Part IV.

Protective friction — Friction that exists on purpose — a second signature, a legal review, a human who can say no — to keep judgment and accountability in the loop, even when it slows things down. Part III. See also MCAS, Part VIII, for what happens when that kind of friction gets removed under commercial pressure instead of redesigned. Protective friction can also fail while kept (see Alarm fatigue), and a check can sometimes be removed safely when the exception it guarded has moved elsewhere (the two-person cockpit, Part VIII).

Reward hacking — When a system finds a way to score well on a task without actually doing what the task was meant to accomplish. A general risk in reinforcement learning, not unique to any one incident. Parts III, IV.

S (Sentido / Meaning) — In the IMPACT formula, the capacity to sustain purpose, belonging, and human connection once economic production requires less of a person's time. It includes a community's

cultural identity, its idiosyncrasy: its own way of being together and of deciding what's worth doing. That is why it is also where the formula holds the part of morality that changes from one culture to another. Part V; central again in Part IX. Meant to be applied as a lens, not calculated as a number — see the opening of Part V.

Social capitalism — The other future Part IX imagines: private ownership and markets continue, but a society accepts that access to income can't depend solely on having a conventional job, and builds mechanisms — a guaranteed income, worker ownership, a social dividend — to distribute the surplus more broadly.

Technostress — The term the psychiatrist Craig Brod coined in 1984, as computers were arriving in offices, for the stress of not being able to live with new computer technologies in a healthy way. Part I.

TMC — Teamwork, Motivation, Communication: the three-part framework built in 2011 to explain why a team with the right strategy can still fail if the people inside it don't trust, understand, or support one another. Part II.

TMC Health — The 2025 revision of TMC: TMC Health = ((T × C × M) / f) × 100, adding friction as the denominator. Part II.

Token — One word, two meanings in this book. In security (Part III), a temporary digital credential that proves who is asking and what they are allowed to do; refreshing it renews it before it expires. In language models (Part IV), the smallest unit of text a model reads and writes, a word or a piece of one.

Transformer — The architecture behind most frontier language models today, introduced in a 2017 paper, "Attention Is All You Need." Its key mechanism is attention. Part IV.

Vertical transmission — Cultural transmission from parent to child: language, habits, half-examined beliefs absorbed before a child can evaluate them. One of three classical channels (Cavalli-Sforza & Feldman, 1981). Part VI.

Waste friction — Friction that serves no real purpose: the meeting that exists because a meeting once existed, the approval chain nobody remembers building. Part III.

XQ (Moral Intelligence) — In the IMPACT formula, the term that asks whether combined human-AI capability is being directed responsibly — toward outcomes that are safe, fair, accountable, and broadly beneficial, rather than simply scaled faster. It doesn't dictate a universal content of the good. It raises questions any culture can answer in its own way and none should skip: who answers, who pays, who benefits, whether it was decided knowingly, whether the people affected knew. Part V. Like S, meant as a lens to apply, not a number to compute.

Sources

By part

Prologue / Part I

- Alvin Toffler, Future Shock (Random House, 1970).
- Carlota Pérez, Technological Revolutions and Financial Capital: The Dynamics of Bubbles and Golden Ages (Edward Elgar, 2002), and her subsequent writing on techno-economic paradigms.
- George M. Beard, American Nervousness: Its Causes and Consequences (Putnam, 1881).
- Craig Brod, Technostress: The Human Cost of the Computer Revolution (Addison-Wesley, 1984).
- American Psychological Association, "2023 Work in America Survey: Artificial Intelligence, Monitoring Technology, and Psychological Well-Being," September 2023, apa.org. Survey conducted by The Harris Poll among 2,515 employed U.S. adults, April 17–27, 2023.

Part III

- OpenAI, "The Hugging Face Incident and the Road Ahead," official incident report, August 26, 2026, openai.com/index/hugging-face-incident-and-the-road-ahead, with the accompanying "OpenAI – Hugging Face Incident Technical Report." Key findings validated by external advisors engaged by OpenAI, including CrowdStrike.
- Dario Amodei, "We Must Pace the Frontier," September 12, 2026, darioamodei.com/post/we-must-pace-the-frontier.
- Sam Altman (@sama) and Elon Musk (@elonmusk), public responses on X to Amodei's essay, September 12, 2026.
- METR and Redwood Research, "Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident," August 26, 2026, metr.org.

- Emily Forlini, "OpenAI, independent firms publish reports on rogue AI attack on Hugging Face. Here are the main takeaways—and what OpenAI still hasn't disclosed," Fortune, August 26, 2026. On the agent counts (about 1,200 agents on the message board, about 700 in the attack).
- "OpenAI saw warning signs weeks before Hugging Face breach," Axios, August 26, 2026. On the technical report's acknowledgment of early warning signs.
- "OpenAI says its AI agents breached its own systems before Hugging Face," Axios, August 6, 2026. On OpenAI's first detailed public account, at Black Hat USA, Las Vegas, August 5, 2026.
- "OpenAI pauses some training amid allegations its rogue agents behaved more badly than first thought," The Register, September 28, 2026. On the second training pause.
- "OpenAI is pausing training of its most powerful AI models after rogue agents hit government sites," Quartz, September 28, 2026. On access to U.S. and Australian government websites.
- "OpenAI abandons plan to release upcoming model as safety concerns escalate," CNBC, September 28, 2026. On GPT-6.1 Astra; first reported by The Wall Street Journal.
- "White House Accord on Super Intelligence: Joint Commitment on Frontier Responsibilities," September 29, 2026, posted by President Trump on Truth Social. Full text in the Washington Examiner.
- "Trump and major AI executives sign 'morally binding' voluntary controls," CBS News, September 29, 2026.
- "OpenAI is sued over rogue AI Hugging Face cyberattack," CNBC, September 30, 2026. On the lawsuit filed by Legal Advocates for Safe Science and Technology.

Part IV

- Ashish Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), 2017.

- Anthropic, "Tracing the Thoughts of a Large Language Model" and the accompanying paper "On the Biology of a Large Language Model," March 27, 2025, anthropic.com/news/tracing-thoughts-language-model — interpretability research on planning in poem generation, cross-lingual "conceptual universality," and parallel arithmetic circuits.
- Adam Tauman Kalai and Ofir Nachum (OpenAI), "Why Language Models Hallucinate," arXiv:2509.04664, September 2025; openai.com/index/why-language-models-hallucinate.
- The author's conversation with a general-purpose AI assistant, October 4, 2026. First-hand account; screenshots and the conversation record are in the author's possession.

Part V

- International Energy Agency, "Energy and AI" (World Energy Outlook Special Report), April 2025, iea.org/reports/energy-and-ai.
- International Energy Agency, "Key Questions on Energy and AI," 2026, iea.org/reports/key-questions-on-energy-and-ai — data-center electricity of 485 TWh in 2025 and about 950 TWh by 2030; technology-company capital expenditure compared with oil and gas investment.
- SemiAnalysis, "$12B of US Ratepayers' Money Wasted on a Modeling Mistake and PJM Wants to Do It Again," August 16, 2026, newsletter.semianalysis.com/p/12b-of-us-ratepayers-money-wasted — analysis of PJM capacity-auction pricing and household electricity costs.
- Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan, "The Moral Machine experiment," Nature 563, no. 7729 (2018): 59–64.

Part VI

- L. L. Cavalli-Sforza and M. W. Feldman, Cultural Transmission and Evolution: A Quantitative Approach (Princeton University Press, 1981).
- "ChatGPT decreases idea diversity in brainstorming," Nature Human Behaviour, May 14, 2025 — a Matters Arising analysis of S. Lee and J. Chung's original 2024 study.

Part VII

- Mata v. Avianca, Inc., No. 1:22-cv-01461 (S.D.N.Y.), sanctions order, June 22, 2023.
- The Nobel Prize in Chemistry 2024, awarded to David Baker, Demis Hassabis, and John M. Jumper — press materials, the Royal Swedish Academy of Sciences.
- AlphaFold Protein Structure Database, EMBL's European Bioinformatics Institute and Google DeepMind.

Part VIII

- "Nationwide real-world implementation of AI for cancer detection in population-based mammography screening" (the PRAIM study), Nature Medicine, January 2025.
- Moffatt v. Air Canada, 2024 BCCRT 149, British Columbia Civil Resolution Tribunal, February 14, 2024.
- Reporting on CNET's AI-written personal-finance articles and subsequent corrections, Futurism, January 2023.
- SAG-AFTRA, 2023 TV/Theatrical Contracts, artificial intelligence provisions.
- U.S. Department of Justice, deferred prosecution agreement with the Boeing Company, January 7, 2021.
- Reporting on Boeing's rejection of a synthetic-airspeed safety system and related whistleblower accounts, The Seattle Times, 2019.

- MIT Sloan School of Management, "How Costco's obsession with culture drove success," Ideas Made to Matter, mitsloan.mit.edu — Jim Sinegal on wages (seventy cents of every dollar spent) and turnover.
- Christine Sinsky et al., "Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties," Annals of Internal Medicine, September 2016.
- Tait D. Shanafelt et al., "Relationship Between Clerical Burden and Characteristics of the Electronic Environment With Physician Burnout and Professional Satisfaction," Mayo Clinic Proceedings 91, no. 7 (2016): 836–848.
- The Joint Commission, "Medical device alarm safety in hospitals," Sentinel Event Alert, Issue 50, April 2013.
- Report of the President's Task Force on Aircraft Crew Complement, July 2, 1981 (available through the Defense Technical Information Center, ADA106889).

Part IX

- Seneca, De brevitate vitae (On the Shortness of Life), 1.3: "Non exiguum temporis habemus, sed multum perdidimus." Translation by the author.
- Tommy Dean Story, "El hilo del crecimiento: de la sabana a la inteligencia artificial" and "¿Hacia una sociedad menos capitalista y más social?" (unpublished essays).
- Associació de Sant Jordi, Alcoy, asjordi.org — the Moros y Cristianos festival, its 28 filaes, and its status as a Festival of International Tourist Interest (1980); 750th anniversary of the patronage of Sant Jordi, 2026.
- Carlota Pérez, as cited under Part I.
- David Autor, "Why Are There Still So Many Jobs? The History and Future of Workplace Automation," Journal of Economic Perspectives 29, no. 3 (2015): 3–30.
- David Autor, "Applying AI to Rebuild Middle Class Jobs," NBER Working Paper No. 32140, February 2024, nber.org/papers/w32140.

- Finnish Government / Kela, "Results of the Basic Income Experiment: Small Employment Effects, Better Perceived Economic Security and Mental Wellbeing," final report, February 2020.
- Alaska Permanent Fund Corporation, Permanent Fund Dividend historical timeline, pfd.alaska.gov/division-info/historical-timeline.