Capability Matters

A CASEBOOK

Learning Engineering for Next-Generation Systems

Capability *Matters*

EDITED BY

William Gray-Roncal, PhD
James Diamond, PhD

JOHNS HOPKINS UNIVERSITY

LEARNING DESIGN AND TECHNOLOGY

EDITION · 15 JUNE 2026

An ongoing artifact of the LENS Practice Flywheel — Identify, Activate, Prototype, Analyze, Transition.

DEDICATION

To everyone who believes in a dream and refuses to give up.

COLOPHON

Learning Engineering for Next-Generation Systems: Capability Matters
A Casebook.
First edition, 2026.
Edited by William Gray-Roncal, PhD and James Diamond, PhD.
Compiled for the LENS specialization within the Learning Design and Technology program at the Johns Hopkins University School of Education.
Set in Instrument Serif and DM Sans.
Printed via Lulu at 8 × 10 in, perfect-bound.

HOW THIS VOLUME WAS MADE

This casebook was assembled using AI tools as part of an iterative learning-engineering process — the same Practice Flywheel (Identify → Activate → Prototype → Analyze → Transition) that the volume's case studies argue for. The methodology treats AI as the dual entity the curriculum names it: a creative partner that accelerated drafting, cross-referencing, layout, and citation lookup, and an epistemic risk that had to be hand-checked against the record. Every case in the book was reviewed by the editors and hand-checked by students for accuracy: confirming that the sources cited exist and are findable, that the quoted attributions are fairly represented, and that the case studies are accurate accounts of the incidents and the investigations they draw on. Items where the source could not be confirmed are marked "Paraphrasing…" so the attribution is honest about what is the author's reconstruction and what is verbatim from the record.

The volume is not finished. It is the first iteration of a reference dataset that LENS will continue to develop. We will revise the book based on reader feedback, reviewer findings, and our own ongoing practice of learning engineering — adding new cases as the operational record grows, correcting our record where it can be improved, and updating the curriculum crosswalk as the program itself iterates. Corrections, new cases, and reader feedback are welcomed.

SOURCES, ATTRIBUTION & LEGAL NOTICES

All cases reference publicly reported incidents and published investigations. Quoted material is attributed to source investigations, court documents, and academic publications. Reflection questions are original to this volume. The audit record of citations checked and changes made is maintained alongside the source repository.

References to convictions, settlements, and findings reflect the public record at the time of writing. Where individuals are named, they are named only as participants in investigations, court proceedings, or public inquiries already in the public domain. The casebook makes no legal characterization beyond what those records establish.

Quoted material is reproduced under fair use (17 U.S.C. § 107) for the purposes of criticism, comment, teaching, scholarship, and research. Trademarks and product names are used nominatively to identify the products and organizations under discussion; no affiliation or endorsement is claimed or implied.

Corrections, requests for clarification, and notice of any inaccuracy can be sent to LDT / LENS at the Johns Hopkins University School of Education. Substantive corrections will be incorporated in subsequent printings and noted in the casebook's public errata.

WELCOME

Learning engineering is a toolbox, not a single tool.

At its center it is a dialogue between two traditions — the engineering disciplines that design, build, and sustain complex systems, and the learning sciences and education that study how people come to know and do. Around that dialogue gather many other fields: cognitive and human-factors engineering, measurement, implementation science, design, data and computation, and the operational domains themselves. The work is plural by design — many disciplines at once — and it draws on domain knowledge, coding, theory, systems thinking, design, teaching, and analysis.

It welcomes everyone who brings a real tool to it. It is not, for that, a free-for-all: learning engineering stands in an intellectual tradition with its own methods, evidence standards, and lineage, and it holds the work accountable to them. No one carries every tool. This is a team sport — capability gets built at the seams between people who each see part of the system clearly and trust the others to see the rest.

So you don't need to know everything before you start. You need to know what you bring, stay honest about what you don't, and find the people who carry the tools you're missing. What you bring is yours to decide. The cases that follow are an invitation to find your part in the work.

— THE EDITORS

INTRODUCTION

Capability is a system parameter.

Every complex system depends on what people can do inside it. That dependency is measurable, designable, and too important to leave to chance.

01. Systems don't fail because the technology breaks. They fail because someone couldn't do what the system required.

02. We engineer every parameter of a system except the people.

03. Capability is not a soft problem. It is a systems engineering problem.

04. Decision-grade evidence. Not opinions about training.

05. Capability without agency is automation. The goal is operators who can perform, adapt, and lead — not comply.

06. Sometimes the constraints make the goal unreachable. Saying so early is a result, not a failure.

The cases in this book are not, individually, hard to understand. A pilot recovered from a stall the wrong way. A nurse could not stop a doctor who skipped a step. A safety operator was watching television on her phone when a pedestrian crossed the road. A captain shot down a civilian airliner because the radar data and the operational framing pointed in opposite directions, and the framing won. The proximate cause is usually obvious. The first investigator on the scene almost always finds it within hours.

What is hard to understand — and what this book exists to make legible — is the system that produced the gap into which each incident fell. In every case the proximate actor was operating inside an architecture of training, procedure, authority, measurement, and incentive that had, for years, been quietly degrading. The pilot had never trained to recover a stall at cruise altitude. The nurse had no procedural path to intervene. The safety operator had been hired into a role nobody had figured out how to make performable. The radar operator had been trained to a tempo and a presumption that the system above him no longer shared with the system around him.

Each of these architectures was designed. Each was funded. Each was reviewed. Each carried, written somewhere in its specifications, the assumption that the people inside it would be able to do what the system required of them when the system required it. That assumption is what this book calls the **capability parameter**. It is a property of the entire sociotechnical system, not of any individual operator. Capability lives at the **interface** between what a system requires of its operators and the impact the system has to deliver — and when that interface is wrong, the gap may originate in

human development, in system design, in organizational performance, or in the interaction among them. Distinguishing which is itself a measurement and engineering problem; the casebook calls it **gap attribution**, and treats it as the diagnostic move that the rest of the discipline depends on. When the interface fails — wherever the source — the consequences are paid in lives, in dollars, in trust, and in time the institution will never recover.

The premise of the Johns Hopkins School of Education's Learning Design & Technology program — and of the Learning Engineering for Next-Generation Systems specialization that has grown out of it — is that the capability parameter is engineerable. Not by accident. Not by declaration. By the same kind of evidence-grounded, methods-grounded, cross-domain discipline that built modern reliability engineering, modern epidemiology, and modern systems safety. The discipline does not yet exist at the scale the evidence demands. LDT and LENS are organized to help build it.

I · THE COST OF THE GAP

The Institute of Medicine's **To Err Is Human** (1999) estimated that preventable medical error caused between forty-four thousand and ninety-eight thousand deaths a year in the United States — already, at the lower bound, a top-ten cause of death. The report catalyzed a generation of patient-safety work: surgical checklists, central-line bundles, team training, computerized order entry. Some of those interventions are documented in this book's What Works chapters as successes. And yet in 2016 Makary and Daniel returned to the question from the same institution that produces this casebook and concluded that medical error remains the third leading cause of death in the United States, killing more than two hundred fifty thousand people a year [1].

The interval between the IOM report and the Makary update is approximately the canonical gap implementation science has identified between research evidence and clinical practice: seventeen years [2]. The average. The median. Only about fourteen percent of research findings ever reach practice at all [3]. The system to surface them, vet them, deploy them, sustain them, and measure their effect at the scale at which they would matter has never been built as an institution. It has been built as a series of grants. The difference is the point.

This is not only a healthcare problem. The U.S. Navy lost seventeen sailors in two avoidable destroyer collisions in 2017 because in 2003 it had cut a sixteen-week classroom-and-simulator course down to a set of CD-ROMs [4]. The world's most expensive aircraft program is operating at half of its design readiness in 2026 because the platform was fielded faster than the maintainer pipeline could be built [5]. One hundred million dollars of educational infrastructure dissolved in fourteen months because the institution building it did not engineer the governance and trust that the technology presupposed [6]. A predictive grading algorithm downgraded a third of an entire nation's high-school students in a single afternoon, and was withdrawn within a week, because nobody inside the agency could be specifically held to account for what the measurement instrument was measuring [7].

None of these failures is a failure of effort. None is a failure of intelligence. Each is a failure of the capability infrastructure to match the system it was operating. In each case, the gap was diagnosable in advance. In most cases, it was actually diagnosed in advance, by someone — and the diagnosis did not produce the remediation. The cost of the unrepaired diagnosis is why **capability** matters.

II · WHAT AN ENGINEERABLE DISCIPLINE LOOKS LIKE

Crew Resource Management did not exist as a discipline in 1976. In March 1977 two 747s collided on a foggy runway at Tenerife and 583 people died. Within five years United Airlines had operationalized the first CRM curriculum [8] within twenty years the Commercial Aviation Safety Team had built the closed-loop hazard-identification system that lets the FAA know whether CRM is still working [9]. The result over the next two decades was an eighty-three percent reduction in U.S. commercial aviation fatality risk and a ninety-five percent reduction in fatalities per hundred million passengers [10]. CRM works not because it taught individual airmanship — pilots were already excellent — but because it redesigned the **system of interaction** in the cockpit: the authority gradient, the communication protocol, the standard brief and debrief. The discipline treated team coordination as an engineerable property of the system.

The same pattern shows up in every domain where the cost of failure has been high enough to fund the discipline. After Three Mile Island the U.S. nuclear industry stood up the Institute of Nuclear Power Operations within nine months — before the Kemeny Commission's report was even finalized — because every utility had recognized that an accident at any single plant would affect every operator's license to operate [11]. INPO and the National Academy for Nuclear Training, founded in 1985, have presided over more than four decades of zero significant releases at U.S. commercial reactors. The comparison with the surface Navy, which cut its training during the same era, is the cleanest available test of what happens when capability is engineered versus when it is deferred.

Healthcare has its own version of this arc, but only partially. Peter Pronovost's central-line checklist, paired with the nurses' authority to enforce it, drove bloodstream infections to near zero across 103 Michigan ICUs and saved an estimated fifteen hundred lives in eighteen months [12]. Atul Gawande's nineteen-item surgical safety checklist, paired with three required pauses in the operative timeline, halved the surgical mortality rate across eight hospitals in eight countries [13]. TeamSTEPPS, jointly developed by AHRQ and DoD, translated five decades of high-reliability research from aviation and the military into clinical practice in years rather than decades — because the implementation infrastructure was funded as part of the intervention rather than as an afterthought [14].

These are not stories about individual hospitals or individual ships. They are stories about the difference between a domain that has organized itself to engineer capability and one that has not. The pattern is consistent across all of them: a catastrophe makes the capability gap visible; an institution is built that treats the gap as the institution's responsibility; the institution funds the measurement architecture that lets it know whether the gap is closing; and twenty years later the death rate is half what it was. The intervention is always paired. A technical artifact — a checklist, a brief, a cord — combined with a cultural artifact — protected authority to use it, a no-blame protocol, a credible governance body. Neither alone moves the curve. Both together move it dramatically.

III · THE DISCIPLINE WE HAVE, THE DISCIPLINE WE NEED

The intellectual material to build this discipline already exists. The learning sciences have produced four decades of converging evidence on how skill, knowledge, and judgment develop and decay [15]. Cognitive engineering has produced rigorous methods for analyzing human-machine systems [16]. Human factors and ergonomics have produced an evidence base for interface and workflow design [17]. Implementation science has produced frameworks for taking effective interventions to scale [18].

Systems safety engineering has produced both diagnostic tools — Reason's swiss-cheese model, Rasmussen's migration-to-the-boundary, Leveson's STAMP — and control-theoretic methods for design [19]. High-reliability- organization theory has documented what the institutions that have solved the capability problem actually do [20]. The discipline is not missing its evidence base. It is missing its integration.

Learning engineering is the integration. The term, in its modern usage, traces to Bror Saxberg and was elaborated by Goodell, Kolodner, and the IEEE Industry Connections Industry Consortium on Learning Engineering (ICICLE) [21]. The IEEE Learning Engineering Body of Knowledge (LEBoK) and its associated process model are the most developed available specification of what the discipline does [22]. The premise is that the work of building, deploying, and sustaining capability at scale is an engineering activity: it requires explicit requirements, evidence-based design, instrumented measurement, and feedback-driven iteration. It is not exclusively a research activity. It is not exclusively a managerial activity. It is engineering, and it can be taught and practiced as engineering.

LENS — Learning Engineering for Next-Generation Systems — is a specialization within the Johns Hopkins School of Education's Learning Design & Technology program that operationalizes this premise for high-consequence operational domains. Its core competencies — capability analysis and requirements, evidence and analytics, human-AI teaming, knowledge transfer, bias and governance, computational and AI methods — are organized to produce graduates who can do the work the cases in this book required, and who can build the institutions that the book's success cases had to invent. The curriculum threading commitments — implementation science as a through-line, equity as a design commitment rather than an audit, decision-grade evidence as a deliverable rather than a report, communication and system-of-systems integration as engineerable properties of the system, the disciplined use of learning- technology aids (from XR to adaptive platforms) as instruments whose evaluation is part of the work, and **AI as both a creative partner and an epistemic risk** — leveraging AI's generative reach while guarding against offloading the thinking — are drawn directly from the patterns the cases reveal.

Three of those commitments deserve a moment of their own — they are the substrate the rest of the curriculum runs on. The first is **communication, translation, and integration within and across disparate systems**. A striking number of cases in this book are, at their proximate cause, translation or integration failures across boundaries — failures of language, convention, unit, role, agency, or discipline to carry meaning reliably from one part of a system to another, or failures of subsystems built to different assumptions to compose into a working whole. Mars Climate Orbiter was a metric-versus- imperial translation failure (Case 98). Tenerife was a translation failure at the takeoff-clearance boundary between two cockpits and a tower under fog and time pressure (Case 117). The Patriot at Dhahran failed because the manufacturer's assumption about run time and the operator's actual run time were on opposite sides of a documentation boundary that had not been engineered to be crossed (Case 129). Eagle Claw failed because the rotary-wing, fixed-wing, ground, and command components were assembled across service boundaries that had no shared operating practice (Case 128). 9/11 was a cross-agency translation failure measured in thousands of dead (Case 135). Boeing 737 MAX was, at one level, an integration failure: a single-string sensor, a flight-control law that assumed pilot mental models from prior models, and a certification chain that did not knit the pieces together (Case 97). The successful cases — AlphaFold (Case 36), MICrONS, CRM (Case 117 paired with Case 118), Keystone (Case 19), the paired-intervention examples in Chapter 8 — are, equivalently, cases of disciplined translation and integration: biology into computation, science into operational practice, technical reform into cultural reform, multiple subsystems into a working

whole. Engineering capability across these boundaries — language, discipline, agency, subsystem-to-subsystem, system-of-systems — is itself a designable target. The LENS curriculum treats it as such.

The second is **technology aids and learning platforms**. The discipline has at its disposal a fast-evolving toolkit: extended- reality (XR) simulation environments — VR, AR, and mixed-reality systems used for procedural training, spatial-reasoning practice, and high-fidelity rehearsal under conditions that would be unsafe or impossible in the field — together with learning- management systems, adaptive learning platforms (the lineage from Cognitive Tutor and ASSISTments to current ITS work; Cases 67, 139), intelligent tutoring frameworks such as GIFT, learning- analytics infrastructures such as xAPI and the Total Learning Architecture (Case 142), game-based learning environments, LLM-augmented tutors and authoring tools, and high-fidelity simulators in aviation, surgery, defense, and process industries. These tools are not the discipline. They are the discipline's instruments — and the cases in this book show, repeatedly, that powerful instruments do not by themselves engineer capability (Cases 53, 54, 2, 139, 22, 157). The LENS curriculum teaches the practitioner to choose, configure, evaluate, and govern these tools with the same evidentiary discipline applied to any other intervention, and to recognize when the binding constraint is the surrounding architecture rather than the tool itself.

The third — and the design constraint on all the others — is **human agency**. The capability discipline can be misread as optimization: produce operators whose behavior matches a specification. That reading collapses the very property the cases reveal as load-bearing. Every successful institution in the What Works chapters — INPO operators trained to challenge their reactor team, Keystone ICU nurses authorized to halt a procedure (Case 19), the Nuclear Navy's questioning attitude (Case 137), the first-officer authority CRM built into the cockpit (Case 117) — engineered for capability **and** for the operator's capacity to exercise independent judgment, to adapt to conditions the system was not designed for, and to shape the surrounding architecture rather than only execute inside it. Capability without that capacity is automation. LENS treats the preservation and development of agency — the room the system leaves its operators to think, decide, and lead — as a design constraint on every intervention, including those in which AI is the partner. The paired risk runs the same direction: an AI collaborator can accelerate the flywheel or quietly offload the thinking, and which one it does is a design and governance decision, not a property of the tool.

IV · THE METHOD · UNPACKING THE CASE

Learning engineering is not a one-time act of design. Its defining feature, as articulated in the IEEE Industry Connections Industry Consortium on Learning Engineering's process model, is iteration: understand the operational context, specify capability requirements, prototype an intervention, instrument it, measure, learn, redesign. The work cycles. The cases in this book that produced sustained outcomes — CRM and CAST, INPO, Keystone ICU, the WHO Surgical Safety Checklist, TeamSTEPPS — all share that loop structure. They were not built once and left alone. They were built, measured, rebuilt, remeasured, and rebuilt again. The institutional capability to sustain the loop **is** what distinguishes them from interventions whose effects decayed.

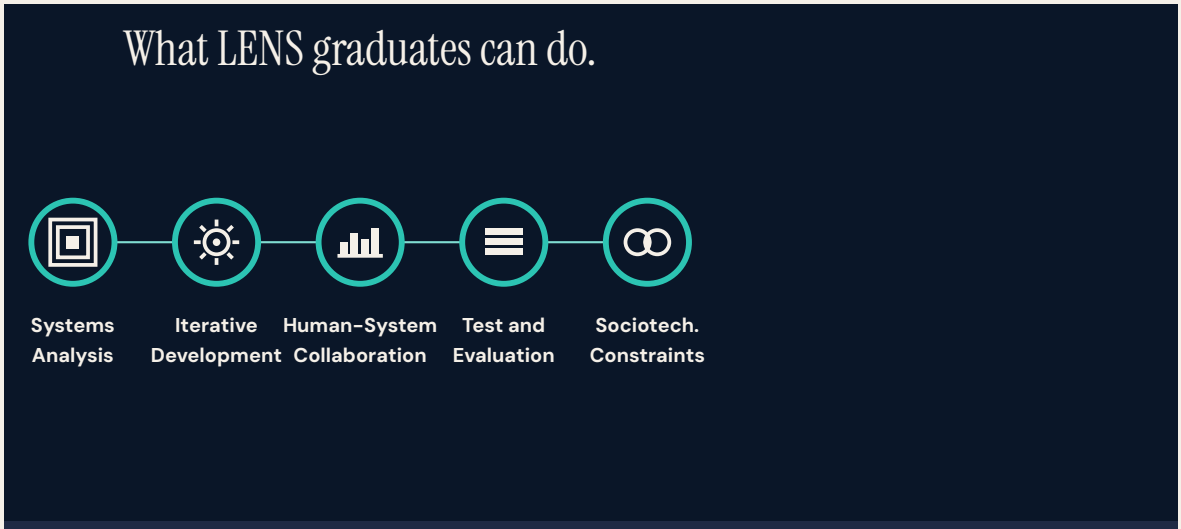
The loop also has a legitimate negative result, and the discipline has to be honest enough to reach it. Capability engineering always operates inside constraints — budget, schedule, regulation, institutional will, politics, ethics, public trust — and many of the binding ones are not technical. Part of the work is **managing those constraints and the risk they carry**: reading the constraint space accurately

enough to recognize when external, non-technical factors have narrowed the solution space to the point where a particular capability goal is no longer realistic to pursue. Reaching that conclusion — and documenting **why** — is itself a valid outcome of a project, not a failure of one. It redirects effort toward goals that are actually reachable, surfaces the non-technical barriers for the stakeholders who can change them, and prevents the far more expensive failure of driving an infeasible target all the way to the operational record before it collapses there. A number of cases in this book are, in part, stories of institutions that could not say "not like this, not yet" in time.

The casebook turns this loop into a reading method. Each case is read, diagnosed, paired with a learning-engineering response, and tested in studio against students' own operational domains. The cycle — framing the system whose capability is at stake, eliciting requirements from operational analysis, building the evidence architecture that lets an institution know whether its interventions worked, examining the human and AI roles inside the designed system, and running the loop end-to-end on a real problem — is the same discipline the casebook applies to every case it reads. The cases that return as success stories in the What Works chapters are the cases whose institutions completed the loop. The What Fails chapters hold the cases whose institutions did not — or did not loop fast enough. Unpacking the case **is** the method: identify, instrument, and close the next such loop before the next case is written.

V · THE ANALYTIC LENS · THE FIVE CAPABILITIES THE CASES TURN ON

The cases in this book turn on five capabilities. They are the analytic lens the casebook reads each case through, and the cases that follow are tagged to them: where capability was absent, one or more of these five was the capability that was missing; where it was engineered, one or more of these five was the capability that was built. Named as concentration learning outcomes, they are also what the discipline sets out to teach.



WHAT LENS GRADUATES CAN DO — FIVE COMPETENCIES

1 · SYSTEMS ANALYSIS — SEE THE WHOLE SYSTEM

SYSTEMS THINKING AND ANALYSIS

Decompose system performance requirements into measurable human capability requirements; apply systems-engineering lifecycle models to scope, sequence, and evaluate capability interventions **within and across disparate systems**; analyze and communicate the interdependencies between human performance and system performance to predict operational impact at scale.

2 · ITERATIVE DEVELOPMENT — BUILD, TEST, REFINE

LEARNING ENGINEERING DESIGN AND IMPLEMENTATION

Design capability-development interventions that integrate learning-sciences principles with measurable outcomes and system-design constraints, and that function at **speed and scale** in operational environments; construct and communicate implementation plans that address adoption, sustainment, and lifecycle integration across organizational and system boundaries.

3 · HUMAN-SYSTEM COLLABORATION — WORK TOGETHER

HUMAN-SYSTEM COLLABORATION AND ADAPTIVE SYSTEMS

Design role architectures, interface and alert systems, mode/state transparency, authority gradients, and recoverability mechanisms that engineer collaborative capability at the human-system boundary — including human-AI partnership as one sub-pattern; evaluate the measured impact of system-mediated work on human performance, naming automation bias, cognitive offloading, and skill atrophy; specify **delegation with revocation** — the disconfirming evidence that would revoke a delegation — and measure collaboration as the unit of analysis.

4 · TEST AND EVALUATION — SHOW WHAT WORKS

DATA, MEASUREMENT, AND EVALUATION

Design ethical instrumentation strategies that produce measurable evidence in regulated or high-consequence environments; construct **decision-grade evidence** artifacts — dashboards, reports, governance documentation — under irreducible uncertainty, naming what is known, what is assumed, and what would change the decision; differentiate capability gaps from system-design and organizational-performance problems using **gap attribution**; demonstrate that omitting a protected attribute does not establish fairness.

5 · NAVIGATING SOCIOTECHNICAL CONSTRAINTS — MAKE IT WORK IN THE REAL WORLD

CONTEXT AND DOMAIN FLUENCY

Analyze the regulatory, organizational, and cultural constraints that shape capability development in healthcare, defense, or education contexts; apply human-systems-integration frameworks to evaluate the measurable impact of capability approaches on operational environments; synthesize and communicate stakeholder requirements across disciplinary and institutional boundaries into coherent design specifications, including cross-regime / platform-dependency

> governance where capability is deployed on a platform governed by a different regime than the one operating it.

See the whole system. Build, test, refine. Work together. Show what works. Make it work in the real world.

Three phrases recur across those five capabilities and across the cases that follow: **within and across disparate systems** (the scope at which capability is actually engineered), **speed and scale** (the operational tempo the discipline must match), and **gap attribution between learning, system design, and organizational performance problems** (the diagnostic that decides what kind of intervention will actually move the outcome). They are the working language the cases are read in.

VI · HOW TO READ THIS BOOK

The book is organized topically, in seven parts: Healthcare & Patient Safety; Education, Training & the Learning Workforce; Aviation & Aerospace; Defense & National Security; Industry, Energy & Enterprise Systems; Disaster Prevention & Recovery; and Algorithms, Governance & Public Systems. Each part splits into two chapters. **What Fails** collects the part's failure cases; **What Works — and the Frontier** collects the interventions that moved the curve and the frontier cases where the discipline is still being asked to specify what good looks like before the catastrophe arrives.

Two analytic threads run across the topical parts. The first is the failure-mode taxonomy carried on every case header: training gap, capability designed out, normalization of deviance, interface failure, governance failure, and knowledge loss. The taxonomy is not a theory. It is a finding: six categories that account for almost every well-documented case in the literature of preventable failure in complex sociotechnical systems, and that recur across every part of this book. The second is the **paired intervention**. Every success case in the dataset shares it: a technical artifact paired with a cultural authority; a measurement instrument paired with a governance body that listens to it; a curriculum paired with the institutional architecture to sustain it. Neither half alone moves the curve. The pair is irreducible. This is the strongest empirical pattern in the book and it directly informs the LENS pedagogical commitment to co-optimization across technical and human design.

One part departs from the domain logic. Part VI, **Disaster Prevention & Recovery**, reads its cases along the lifecycle of a disaster instead: before the event — the prevention regimes whose quiet thinning is the real story — and after it, the response, the recovery, and the institutions built (or not built) so the same failure cannot recur. Cases whose load-bearing lesson belongs to a domain stay in their domain parts; what qualifies a case for Part VI is that its lesson is the pre/post lifecycle itself.

Each case occupies a two-page spread. The left page tells what happened — narrative, evidence, attribution. The right page is the **Learning Engineering Lens:** a synthesis of what the case teaches, how the LENS curriculum addresses the pattern, and reflection questions designed for studio discussion. The casebook is built to be taught from, read straight through, or consulted as a reference. The curriculum crosswalk in the closing matter maps each case to the specific courses in the program that use it as a worked example.

How these cases were chosen. The collection is purposive, not systematic. The cases are exemplars — drawn from the documented public record and from the editors' own practice — selected because each makes a capability-engineering pattern legible, not because they constitute a representative or exhaustive sample. The corpus over-represents catastrophic, well-investigated failures, which generate the richest public record, and under-represents the everyday, program-scale work that fills most of a practitioner's career; a deliberate expansion has begun to close that gap. Read the book as a teaching instrument rather than a systematic review: the claim is that these patterns recur and are engineerable — not that this is every case, or the average one.

HOW TO USE THIS BOOK

A field manual for capability engineering.

This book collects a growing record of real incidents in which capability — what people could or could not do inside a complex system — determined whether the system worked. Some are failures: lives lost, systems wrecked, billions written off. Some are successes: deliberate interventions that engineered capability into the architecture of the work. Together they form an evidence base for the argument that capability is a designable, measurable property of every complex system.

Each case occupies a two-page spread. The left-hand page tells the story: what happened, what investigators found, and the documentary record. The right-hand page is the **Learning Engineering Lens** — what the case teaches about capability as a system parameter, how the LENS curriculum addresses the pattern, and a short set of reflection questions designed for studio discussion.

The Lens page also carries a **Who Builds This** note: the mix of expertise and tools a team would bring to engineer the fix, read off the case's failure modes. A Training-Gap case pulls in learning scientists and instructional designers; an Interface case, human- factors and interaction designers; a Governance case, policy, ethics, and measurement. Most cases need several at once, held together by domain experts and a learning engineer. No single discipline carries a case alone — reading the modes as a staffing list is itself a LENS skill, and what **you** bring is yours to decide.

READING THE TAXONOMY

Each case is tagged with one or more failure modes. Most cases involve several. The primary mode is listed first; contributing modes follow.

T Training Gap
Personnel were never taught what they needed to know.

D Designed Out
Human capability was deliberately removed from the system.

N Normalization of Deviance
Known risks were accepted as routine.

H Human-System Interface
Correct performance was made unreasonably difficult.

G Governance & Trust
Ethics, accountability, and stakeholder trust were afterthoughts.

K Knowledge & Institutional Memory
Organizational knowledge degraded or was never captured.

LENS COURSE CODES

Each case maps to one or more courses in the LENS specialization:

SHARED LDT FOUNDATIONS · ADDITIONAL CONTEXT

F1 **Learning Sciences Studio.** How learning, motivation, and skill develop — applied to technology-mediated design.

F2 **Critical Perspectives on Educational Technology.** How power, equity, and society shape — and are shaped by — learning tech.

LENS-DESIGNED COURSES

LEN 1 Principles of LE for Complex Systems

LEN 2 Human-AI Teaming and Adaptive Systems

LEN 3 Learning Engineering Systems

LEN 4 Evidence, Analytics, and Measurement

LEN 5 Human Capability Analysis and Requirements

LEN 6 Applied Problem Solving in LE

LEN 7 Bias, Risk, and Governance

LEN 8 Knowledge Transfer and Organizational Learning
LEN 9 Computational and AI Methods
LEN 10 Learning Engineering Project (capstone)

> *The strongest cases pair a catastrophic failure with a systematic capability intervention. Together they show both the cost of the gap and the tractability of the solution. The discipline that closes that gap is learning engineering.*

THE CASE MATRIX

The cases at a glance

Gold numbers indicate failures and systemic conditions; teal numbers indicate paired-intervention successes and the open closing case. Numbers marked ° appear in the complete digital edition but not the printed main volume; unmarked cases appear in both. Numbering is global across the whole corpus.

1	Therac-25 1985 – 1987	H·D·G
2°	EHR / CPOE Implementation 2005 – present	H·D·G
3°	Watson for Oncology — Delegation Marketed Ahead of Capability 2013 – 2018	D·K·N
4°	Mid Staffordshire NHS Foundation Trust 2005 – 2009	G·N·K
5	Epic Sepsis Model 2017 – 2021	D·G·N
6	Racial Bias in Pain Assessment — The False-Belief Mechanism 2016	T·K·N
7°	VA Wait-Time Scandal 2014	G·K·N
8°	Medical Errors as Systemic Failure 1999 – present	T·H·N·K·G
9°	Vioxx Withdrawal 1999 – 2004	G·D
11°	California Nurse Staffing Ratios — Legislating a Capability Requirement 2004 – 2010	G·K
12°	ACGME 80-Hour Resident Duty-Hour Reform 2003–2017	T·K·N
13°	Implementation Science in Healthcare — The 17-Year Gap ongoing	K·G·N
14°	Barsuk SBML — Simulation-Based Mastery Learning Dissemination from Northwestern to the VA 2009 – 2020s	T·K
15°	I-PASS Handoff Bundle — Structuring the Human-to-Human Transfer 2014	H·K·N
16°	NCSBN National Simulation Study — Licensing the 50% Substitution Rule 2014	G·K·D
17°	Spaced Education RCTs in Medical Training 2007 – 2009	H·K·D
18°	PEPFAR HIV Training Across 16 Sub-Saharan African Countries — Modality Comparison Under Disruption 2023	K·N·D
19	Keystone ICU / Pronovost Checklist 2004 – present	T·N
20	TREWS / Sepsis Watch 2018 – 2022	H·K·G
21°	Sterile-Cockpit Ward Rounds — Adapting an Aviation Principle to Clinical Handoff 2024 – 2025	H·K·N
22°	ChatGPT in Healthcare — Hallucination Cases 2023 – present	H·D
23°	WHO Surgical Safety Checklist 2008 – present	T·N

Modes. T Training Gap · D Designed Out · N Normalization of Deviance · H Human-System Interface · G Governance & Trust · K Knowledge & Institutional Memory

Courses. The LDT / LENS courses each case serves as a worked example for are indexed separately in **The cases by LENS course** in the appendix.

Parts. The book is organized topically in seven parts — Healthcare & Patient Safety; Education, Training & the Learning Workforce; Aviation & Aerospace; Defense & National Security; Industry, Energy & Enterprise Systems; Disaster Prevention & Recovery; Algorithms, Governance & Public Systems. Each part splits into **What Fails** and **What Works — and the Frontier**; the open question closes the volume at Case 205.

Chapter 1

Healthcare & Patient Safety — What Fails

When care systems assume capability that was never specified, trained, or measured.

The harm arrived through the workflow, not around it.

AN OBSERVATION RECURRING ACROSS THE CHAPTER'S CASES.

CASE 1 HEALTHCARE TECHNOLOGY 1985 – 1987

Therac-25

H Human-System Interface D Designed Out G Governance & Trust

IMPACT **Six known massive radiation overdoses; at least three deaths; canonical case in software safety engineering**

IN BRIEF

The Therac-25, a radiation-therapy machine, massively overdosed at least six patients between 1985 and 1987, killing at least three. Its predecessors used hardware interlocks to stop the high-energy beam from firing without its target in place; judging the interlocks not worth duplicating and trusting software over hardware, the Therac-25 removed them and relied on software — adapted from the older machines and never engineered for safety — to keep the beam modes separated. At least two distinct software defects drove the overdoses — a data-entry race condition behind the cryptic "Malfunction 54" deaths at Tyler, and a counter overflow that silently skipped a collimator check in the fatal Yakima accident — either of which could fire the full beam with no target in place. Leveson and Turner's 1993 investigation, the founding case of software-safety engineering, found a systemic failure: a safeguard removed with nothing put in its place. The lesson — a safeguard you delete does not remove the hazard; it relocates it to whatever you failed to build.

BACKGROUND

The Therac-25 was a radiation-therapy linear accelerator. Its predecessors used hardware interlocks that physically blocked the high-energy beam unless the spreading target was confirmed in place — a mechanical backstop that could not be talked out of stopping the beam. To save cost and simplify the machine, the Therac-25 removed them and trusted software — adapted from the older machines and never engineered from the ground up for safety-critical use — to keep its two beam modes, a hundredfold apart in energy, safely separated. The safety case migrated silently from steel to code.[0]

WHAT HAPPENED

Between 1985 and 1987 the machine massively overdosed at least six patients. A race condition the engineers never knew about meant that if an operator entered a prescription, caught a mistake, and corrected it within about eight seconds — the speed an experienced operator naturally reaches — the full-power beam could fire with no target in place, delivering up to a hundred times the intended dose. The console showed only "MALFUNCTION 54," a code with no documented meaning, and offered to proceed; operators, assured the machine was safe and long accustomed to its cryptic faults, did. That fast-edit race condition was only one of two independent defects: the fatal 1987 Yakima overdose came from an unrelated bug — a status counter that overflowed to zero once every 256 setups and silently skipped

a collimator check — evidence the failure was systemic, not a single stray line of code. At least three patients died of the radiation burns.[1]

THE INVESTIGATION

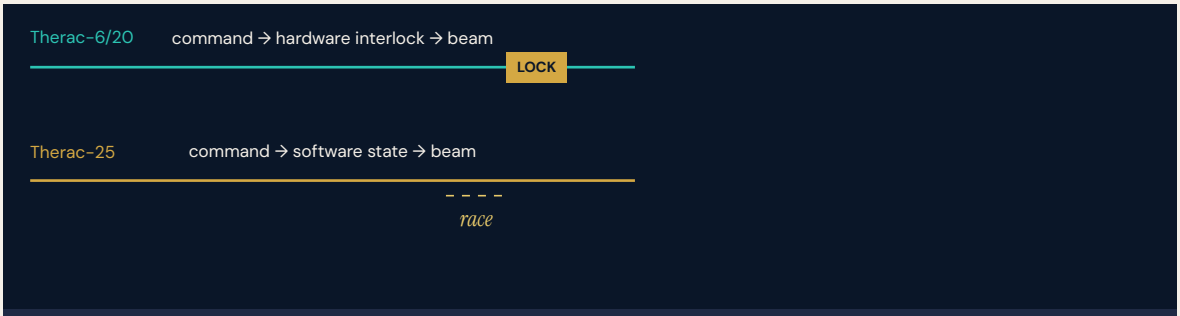
When patients reported searing pain, the manufacturer first insisted an overdose was impossible and treated each report as an isolated complaint rather than a signal. Nancy Leveson and Clark Turner's 1993 investigation — the founding case study of software-safety engineering — found the fault was systemic rather than a single bug: overconfidence in software, removal of the hardware safeguards without replacement, meaningless error messages, reused code never audited for safety, no independent review of the safety-critical logic, and an incident-reporting posture that dismissed the early warnings instead of compounding them into evidence.[2]

THE CAPABILITY GAP

The hardware interlock had not been redundant; it **was** the safety case, the one thing standing between a typing error and a lethal dose. Removing it put nothing in its place — no verified software check, no informed operator action, no independent monitor watching the beam. The operator stayed nominally in the loop but lost any means to detect what the machine was doing wrong, since the only feedback was a code that told them nothing, which made the human presence a formality rather than a safeguard. The question capability engineering would have forced — **what function now carries the interlock's load?** — was never asked of the redesign.[3]

AFTERMATH & REFORM

Therac-25 reshaped safety-critical software practice, making the case by counterexample for independent hazard analysis, safeguards that do not all rest on the same software, error messages that say what is wrong so the operator can act, and treating field reports as evidence to be aggregated rather than complaints to be closed. It remains the canonical teaching case across software, medical-device, and systems-safety curricula.[4] Its lesson is exact and portable: a safeguard you remove does not remove the hazard it guarded — it relocates the hazard to whatever you failed to put in its place, and waits there.



SOFTWARE RACE CONDITION REPLACED HARDWARE INTERLOCK

> *Therac-25 illustrates the dangers of relying on software safety controls without rigorous engineering practices.*
>
> PARAPHRASING LEVESON & TURNER, IEEE COMPUTER, 1993

REFERENCES

1. N. G. Leveson & C. S. Turner, "An Investigation of the Therac-25 Accidents," IEEE Computer 26(7): 18–41 (1993). doi:10.1109/MC.1993.274940 — the removed hardware interlocks and the adapted software.
2. Leveson & Turner (1993) — the two software defects (data-entry race condition and counter overflow), the uninformative "Malfunction 54", overdoses up to 100×, six accidents, and at least three deaths.
3. Leveson & Turner (1993); N. G. Leveson, Safeware: System Safety and Computers (Addison-Wesley, 1995) — the manufacturer's denial and the systemic findings.
4. N. G. Leveson, Engineering a Safer World: Systems Thinking Applied to Safety (MIT Press, 2011) — why removing a safeguard requires explicitly reassigning its safety function.
5. The case's role in medical-device software regulation and software-safety practice (FDA software guidance; IEC 62304 lineage); see also Ethics Unwrapped, UT Austin.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Therac-25 is the moment when removing the human safety margin became visible as a design choice. The hardware interlock was not redundant — it was the safety case. When it was removed, no equivalent capability was put in its place. The operator was retained in the system but stripped of the ability to detect what it was doing wrong. Capability engineering would have asked, before removing the interlock, **what function takes its load?**

LENS APPROACH

LENS frames Therac-25 as a **Design-Out** failure made visible through Interface and Governance pathologies. Studio projects in LEN 5 ask students to produce capability-load diagrams tracing every safety function to a hardware backstop, a software check, or a trained operator action with the information needed to perform it. LEN 7 examines incident reporting as governance.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
HEALTHCARE TECHNOLOGY	→ H D G →	D1 D2 **D3** D4 D5 Human-System Collaboration	3	3.1	3

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Before deleting any hardware interlock, write down the safety function it performs and explicitly reassign that function to a verified software check, an independent monitor, or an informed operator action.
- Do not let safety rest entirely on one software path; require an independent channel that does not share the same code, so a single defect cannot defeat the whole safety case.
- Specify error messages as a safety deliverable — each fault code must say what is wrong and what the operator should do, never offer to proceed past an undiagnosed condition.

IN OPERATION / AFTER

- Treat every field report of unexpected harm as evidence to be aggregated, not a complaint to be closed, with a standing path to halt the device when a pattern emerges.
- Audit any safety-critical code reused from a prior system against the new hazard set, since assumptions safe in the old context may be lethal in the new one.
- Instrument the machine so an independent monitor can detect a beam fired without its target and stop it, restoring the interlock's function even where the operator cannot see the fault.

REFLECTION QUESTIONS

1. Identify a system in your domain that migrated a safety function from hardware to software. Where did the human-capability load go, and who is accountable for sustaining it?

2. Therac-25 operators saw “MALFUNCTION 54” and continued treatment. Redesign that interface using LEN 2 principles so that the operator’s correct action is the easiest action.

3. The Therac-25’s safety-critical software was inherited from earlier machines and never re-audited. Identify reused code in a system you build that now carries a load it was never written for, and specify the review it should have received.

WHO BUILDS THIS human factors & interaction design · systems & human-factors engineering · governance, ethics & measurement — plus domain experts and a learning engineer to integrate. TOOLS usability testing · cognitive walkthroughs · task analysis · design prototyping · governance frameworks · bias & evidence audits

CASE 5 HEALTHCARE CLINICAL AI GOVERNANCE 2017 – 2021

Epic Sepsis Model

D Designed Out G Governance & Trust N Normalization of Deviance

IMPACT **External validation across 38,455 hospitalizations found AUROC 0.63 versus reported 0.76–0.83, missing ~67% of sepsis at the operational threshold; deployed in hundreds of hospitals without independent validation or FDA clearance**

IN BRIEF

The Epic Sepsis Model was a proprietary machine-learning sepsis prediction tool embedded in the Epic EHR and deployed in hundreds of US hospitals — the most widely operational clinical AI in American medicine. Until Wong et al. (JAMA Internal Medicine 2021) ran an external validation across 38,455 hospitalizations, no independent evaluation had been published. The reported AUROC was 0.63, well below the 0.76–0.83 Epic had reported; at the operational threshold, the model missed roughly two-thirds of sepsis cases, with a 12% positive predictive value and substantial alert burden. The case is not principally about the model's performance. It is about the governance seam that let the deployment happen at scale without independent validation: as an EHR-embedded proprietary feature, the model sat outside the FDA's medical-device oversight regime, so the machinery that would have surfaced its limitations at clearance was never engaged. The post-deployment surveillance pattern (Vioxx, Case 9) is the analog: the harm was the absence of a standing system to check the tool once it was in the population's hands.

BACKGROUND

By the late 2010s the Epic Sepsis Model was, by deployment count, the most widely operational clinical AI in American medicine — embedded in hundreds of hospitals as a default feature of the dominant inpatient EHR. The tool was presented as a help to clinicians trying to catch sepsis earlier in a workflow already saturated with alerts. The model's design and validation, however, were proprietary and not externally evaluated.[0]

WHAT HAPPENED

In 2021 Wong et al. published the first large external validation in JAMA Internal Medicine: 38,455 hospitalizations at the University of Michigan, with sepsis diagnosed using two consensus definitions. The model achieved an AUROC of 0.63, materially below Epic's reported 0.76–0.83. At the operational threshold for bedside alerting, the model missed roughly two-thirds of sepsis cases. The 12% positive predictive value meant most alerts were false; the alert burden landed on clinicians who could not distinguish the few real alerts from the many spurious ones.[1]

THE INVESTIGATION

The governance seam is the structural lesson. Because the Epic Sepsis Model was distributed as a feature of an EHR rather than as a stand-alone clinical-decision-support device, it did not require FDA clearance. The machinery that would normally require independent validation, post-market surveillance, and demographic stratification of performance was never engaged. The model's deployment was a regulatory non-event because the regulatory regime treated the EHR layer as out of scope. The clinical AI question and the device-oversight question diverged.[2]

THE CAPABILITY GAP

What surfaced the failure was post-deployment external validation — the exact discipline that the clearance pathway omits. The Wong et al. paper was disconfirmation in the form the system did not otherwise provide. In October 2022 Epic overhauled the model — changing the sepsis-onset definition, reducing its reliance on antibiotic-order signals, and moving to a gradient-boosted model (ESM v2) that hospitals must train on their own local data before use; many had already turned the alert off, recalibrated, or replaced it. Later multicenter validations of the updated model reported stronger discrimination (AUROC roughly 0.82–0.92) but persistent institution-to-institution variability, low positive predictive value, and heavy alert burden — reinforcing that local validation remains necessary. The corrective action worked through publication, not through governance. That is the gap: a tool can be deployed at hundreds of sites, alert at the bedside for years, and still be disconfirmable only by an academic paper rather than by the surveillance architecture the deployment was supposed to have.[3]

AFTERMATH & REFORM

The Epic case is the negative pair to TREWS (Case 20) and the governance-seam analog to Radiology AI Miscalibration (Case 35). Together they teach a typology: delegation done well as a paired intervention with engineered interface and outcome-grounded evidence (TREWS); delegation done badly as a proprietary EHR-embedded model deployed outside the device-oversight regime without independent validation (Epic); delegation halted on rights grounds because the system was both ineffective and rights-violating (SyRI); delegation marketed ahead of capability (Watson for Oncology). The four together are the canonical AI delegation typology.

> *A deployment is not a validation. Deployment without independent validation is delegation without evidence.*
>
> EDITORS' SYNTHESIS OF WONG ET AL. (2021).

REFERENCES

1. Wong et al. (2021), "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients," JAMA Internal Medicine 181(8):1065–1070, doi:10.1001/jamainternmed.2021.2626.
2. Habib et al. (2021), commentary on Wong et al., JAMA Internal Medicine — on the implications for proprietary clinical AI.
3. FDA, Clinical Decision Support Software: Final Guidance (2022) — the post-Wong reframing of the EHR-embedded oversight question.
4. Ross, C. (2022), "Epic overhauls its sepsis algorithm," STAT News (Oct 2022) — the ESM v2 redesign and the shift to locally-trained models.
5. Adams et al. (2022), Nature Medicine — the paired positive case (20).

THE LEARNING ENGINEERING LENS

LE INSIGHT

The Epic Sepsis Model is the canonical case of consequential clinical-AI delegation at scale without independent validation. The structural lesson is not the model's poor performance; it is the governance seam that let the deployment proceed without the validation and surveillance machinery that the medical-device pathway would have required, surfaced only by post-deployment external work.

LENS APPROACH

Epic is the Domain 4 + Domain 3 / Problem Type 6 failure that motivates the post-deployment-surveillance discipline LENS teaches. Used in Domain 4 (Test and Evaluation) for measurement architecture under proprietary opacity and the gap-attribution LEO; in Domain 3 (Human-System Collaboration) for the delegation-with-revocation LEO — Epic was delegated without a pre-specified revocation criterion; and in Domain 5 (Navigating Sociotechnical Constraints) for the cross-regime / platform governance seam. Pairs directly against TREWS (Case 20) and sits in the AI-delegation typology with SyRI and Watson.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
HEALTHCARE CLINICAL AI GOVERNANCE	D G N	D1 D2 D3 **D4** D5 **Test and Evaluation** + D3	6	2.4	4 · 5 · 3

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Require independent external validation before deployment of any consequential clinical AI, regardless of whether it ships as a stand-alone device or as a feature of a host platform.
- Specify in advance the disconfirming evidence — population, threshold, PPV, alert burden — that would revoke the delegation, and the channel through which that evidence would surface.
- Identify the regulatory regime the tool falls under, and where the seam between regimes is — proprietary EHR features should not be exempt from clinical-AI oversight by virtue of their packaging.

IN OPERATION / AFTER

- Build post-deployment surveillance as a standing institutional capability — outcome metrics, demographic stratification, alert-burden audit — so disconfirmation does not require a single academic paper to surface.
- Close the cross-regime seam: clinical AI embedded in EHRs should be subject to the same independent validation and surveillance as stand-alone clinical-decision-support devices.
- When disconfirming evidence arrives, treat it as a designed input: revise, recalibrate, or remove on a defined timeline, with the corrective action visible to the clinicians who used the tool.

REFLECTION QUESTIONS

1. Identify a clinical AI tool deployed in your domain. Where in the regulatory architecture would independent validation have been required, and where could it slip the seam? What pre-specified disconfirming evidence would revoke the delegation?

2. Design the post-deployment surveillance deliverable that should accompany every deployment of consequential clinical AI — including embedded-in-EHR features that currently fall outside the device-oversight regime.

3. The disconfirmation in this case came from a single academic paper, not from a standing institutional architecture. What is the minimum surveillance machinery that would have surfaced the model's performance gap at the operational threshold without requiring the Wong et al. paper to exist?

WHO BUILDS THIS systems & human-factors engineering · governance, ethics & measurement · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS task analysis · design prototyping · governance frameworks · bias & evidence audits · incident analysis · safety dashboards

CASE 6 CLINICAL MEDICINE MEDICAL EDUCATION HEALTH EQUITY 2016

Racial Bias in Pain Assessment — The False-Belief Mechanism

T Training Gap K Knowledge & Institutional Memory N Normalization of Deviance

IMPACT **About half of surveyed white medical students and residents endorsed at least one false belief about biological differences between Black and White people; those who held more false beliefs rated Black patients' pain as less severe and recommended less accurate treatment**

IN BRIEF

Hoffman, Trawalter, Axt, and Oliver (PNAS 2016) surveyed medical students and residents on a battery of false beliefs about biological differences between Black and White people (e.g., "Black people's skin is thicker," "Black people's nerve endings are less sensitive"). About half endorsed at least one such belief. The paper's experimental layer showed that respondents who endorsed more false beliefs rated the pain of mock Black patients as less severe than the same pain in mock White patients, and made less accurate treatment recommendations. The mechanism the case identifies is specific and unusually precise for a bias study: the pain-assessment gap is traceable to a small set of nameable false biological beliefs, not to diffuse implicit bias or structural-only explanation. That precision is what makes Hoffman the human-development case in the race-construct trio. The capability deliverable is not awareness training; it is curriculum that specifically disconfirms the named false beliefs and instrumentation that surfaces when bedside ratings of pain diverge by patient race.

BACKGROUND

Pain assessment is a clinician's judgment, made repeatedly across a day, on patients whose subjective report of pain has to be translated into a numeric rating and a treatment decision. A documented finding in the medical literature is that Black patients in the United States are systematically under-treated for pain across emergency-department, post-surgical, and end-of-life settings. The bias has been measured at the population level for decades; the mechanism was less precisely named.[o]

WHAT HAPPENED

Hoffman et al. (2016) administered a battery of statements about biological differences between Black and White people to 222 white medical students and residents — some true, some false (e.g., "Black people's skin is thicker," "Black people's blood coagulates more quickly," "Black people's nerve endings are less sensitive"). About half of respondents endorsed at least one false belief; a smaller subset endorsed several. The experimental layer

of the study presented respondents with two mock patient cases identical except for race, asked them to rate the patients' pain, and asked them to recommend treatment.[1]

THE INVESTIGATION

The pattern was that respondents who endorsed more false beliefs rated the pain of the Black mock patient as less severe than the pain of the White mock patient, and recommended less accurate treatment for the Black mock patient. Respondents who endorsed fewer false beliefs showed the opposite pattern, rating the Black mock patient's pain as slightly higher. The case is unusual in identifying a specific cognitive mechanism — a small set of named false biological beliefs — that mediates a documented population-level disparity. Most bias studies leave the mechanism diffuse; Hoffman names it precisely enough that a curriculum or assessment can target it.[2]

THE CAPABILITY GAP

What the case teaches at the construct layer is that the capability deliverable in medical education is not generic "implicit bias" awareness — it is curriculum that specifically disconfirms the named false beliefs, with assessment instruments that test whether the beliefs were actually disconfirmed. Operationally, the deliverable is a bedside instrument or surveillance pattern that surfaces when pain ratings diverge by patient race in ways that survive case-mix adjustment. The Hoffman finding makes both deliverables specifiable in a way that more diffuse bias findings did not.[3] The mechanism has since proven portable in a way that extends the deliverable: Omiye et al. (2023, npj Digital Medicine) found that commercial large language models reproduce the very false beliefs Hoffman catalogued — about skin thickness, nerve density, and pain tolerance — when queried with medical prompts, so the disconfirmation task now reaches past the medical-school curriculum to the AI systems entering the same clinical workflows.[4]

AFTERMATH & REFORM

Hoffman pairs with pulse oximetry (Case 26) and eGFR (Case 25) in the race-construct trio. The three cases are the same surface harm — minority patients systematically under-served across a clinical decision — attributable to three distinct layers of the system: the construct definition (eGFR), the validation architecture (pulse oximetry), and the human-development mechanism (Hoffman). The trio is the case-grounded basis for the LEO **Gap attribution**: distinguishing capability gaps attributable to human development, system design, and organizational performance, and selecting measurement that isolates the intended cause.

> *The mechanism the paper names is precise enough to teach against. Awareness training is not a curriculum; a curriculum has to disconfirm something specific.*
>
> EDITORS' SYNTHESIS OF HOFFMAN ET AL. (2016).

REFERENCES

1. Hoffman, Trawalter, Axt, & Oliver (2016), "Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites," PNAS 113(16):4296–4301, doi:10.1073/pnas.1516047113.
2. Anderson, Green, & Payne (2009), "Racial and ethnic disparities in pain: causes and consequences of unequal care," Journal of Pain 10(12):1187–1204 — the population-level disparity.
3. Sabin & Greenwald (2012), "The influence of implicit bias on treatment recommendations for 4 common pediatric condi-
4. Vyas, Eisenstein, & Jones (2020), NEJM — connecting race-in-clinical-algorithms to race-in-clinical-judgment.
5. Omiye, Lester, Spichak, Rotemberg, & Daneshjou (2023), "Large language models propagate race-based medicine," npj Digital Medicine 6:195 — commercial LLMs reproduce the Hoffman false beliefs.

tions," American Journal of Public Health — the diffuse-mechanism backdrop the Hoffman precision improves on.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Hoffman et al. is the human-development case in the race-construct trio. The pain-assessment disparity in medical settings is mediated, in measurable part, by a nameable set of false biological-difference beliefs held by clinicians in training. The capability deliverable is curriculum that specifically disconfirms the beliefs and instrumentation that surfaces the bedside rating gap when it persists.

LENS APPROACH

Hoffman is the human-development case in the race-construct trio (Cases 25, 26 and 6). LENS uses it in Domain 4 (Test and Evaluation) for the LEO **Gap attribution** — the gap is in the clinician's training, not the construct or the device — and in Domain 2 (Learning Engineering Design) for the curriculum-design implication. The trio together is the case-grounded basis for **Gap attribution**: same surface harm, three distinct layers, three distinct remediations. Adjacent to the lending pair (Cases 186–113) at the construct layer.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
CLINICAL MEDICINE MEDICAL EDUCATION HEALTH EQUITY	T K N	D1 D2 D3 **D4** D5 Test and Evaluation	5	8.1	4

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Build curriculum that specifically disconfirms the named false biological beliefs identified in the Hoffman survey, with assessment items that test whether the disconfirmation took hold.
- Instrument bedside pain ratings to surface case-mix-adjusted divergence by patient race; the gap is otherwise invisible to the system that produces it.
- Identify the layer of the gap before designing the remediation: construct, validation, or human-development. A construct fix cannot remediate a clinician-belief fix and vice versa.

IN OPERATION / AFTER

- Re-administer the Hoffman survey periodically as a curriculum-evaluation instrument; a curriculum that does not move the false-belief endorsement rate is not closing the mechanism the paper identifies.
- Track whether the bedside pain-rating disparity narrows in cohorts that received the disconfirming curriculum, with reporting at intervals long enough for selection effects to settle.
- Cross-reference the human-development result against the construct (eGFR) and validation-architecture (pulse oximetry) layers, so the overall equity capability of the clinical system is not assessed only at the layer the institution finds easiest to instrument.

REFLECTION QUESTIONS

1. Identify a documented disparity in your domain whose mechanism is treated as diffuse. What would a Hoffman-style survey look like — a battery of named false beliefs or assumptions whose endorsement could be measured and whose presence predicts the operational decision?

2. Design the curriculum-evaluation instrument you would use to test whether a curriculum has actually disconfirmed the false beliefs. What endorsement-rate change would you require before claiming the mechanism has been addressed?

3. Hoffman is the human-development case in the race-construct trio. Pulse oximetry (Case 26) is the validation-architecture case; eGFR (Case 25) is the construct-definition case. Which of the three layers does your domain currently address, and which does it leave invisible?

WHO BUILDS THIS learning science & instructional design · knowledge management & data engineering · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS scenario simulation · competency models · knowledge bases · data pipelines · incident analysis · safety dashboards

Chapter 2

Healthcare & Patient Safety — What Works — and the Frontier

When measurement, teamwork, and design close the loop on patient harm.

Every durable safety gain shipped with an owner and an instrument.

AN OBSERVATION RECURRING ACROSS THE CHAPTER'S CASES.

CASE 19 HEALTHCARE 2004 – PRESENT

Keystone ICU / Pronovost Checklist

T Training Gap N Normalization of Deviance

DISCLOSURE Institutional overlap: an editor shares an institution (Johns Hopkins) with the study's lead author; no editor was personally involved in the Keystone ICU work. Included on the published peer-reviewed evidence (NEJM 2006) and its sustained, independently replicated results.

IMPACT **Central-line-associated bloodstream infections (CLABSI) reduced to near zero across 103 Michigan ICUs; ~1,500 lives saved in 18 months; ~$175M saved; sustained at ten years**

IN BRIEF

Peter Pronovost's central-line checklist has five items — wash hands, use full barrier precautions, clean the skin with chlorhexidine, avoid the femoral insertion site, and remove unnecessary catheters — and not one of them was unknown to the physicians it governed. The question was never what to do, but whether it would be done every time. The Keystone project, launched across 103 Michigan ICUs in 2004, paired the checklist with a cultural change: nurses were not merely permitted but required to stop the procedure if a step was skipped. That authorization is the element the case treats as load-bearing — though it rode inside a multi-component bundle (clinician education, a dedicated line cart, daily catheter-review goals, and monthly infection-rate feedback) that the trial tested as a whole and could not decompose. Central-line-associated bloodstream infections fell to near zero, an estimated 1,500 lives and $175 million were saved in eighteen months, and the effect was sustained at ten years.

BACKGROUND

Central-line-associated bloodstream infections were a common, often fatal complication of intensive care, and the steps to prevent them were well established and uncontroversial. The problem was reliability: in the existing culture, a nurse who saw a physician skip a sterile step had no procedural path to intervene without crossing the hospital's authority gradient. The knowledge was universal and the steps cheap; what was missing was any mechanism that made the right action happen every time rather than most of the time, which is precisely where a fatal infection finds its opening.[0]

THE INTERVENTION

In 2004, Peter Pronovost's team launched the Keystone ICU project across 103 Michigan units. It combined a simple five-item central-line checklist — hand hygiene, chlorhexidine skin prep, full-barrier draping, sterile gown-mask-gloves, and a sterile dressing — with an explicit authorization: nurses were required, not merely permitted, to stop any procedure in which a step was missed. The distinction between permitted and required was deliberate

— a permission a nurse might decline to exercise against a senior physician became an obligation the institution stood behind, removing the personal risk of intervening.[1]

HOW IT WORKED

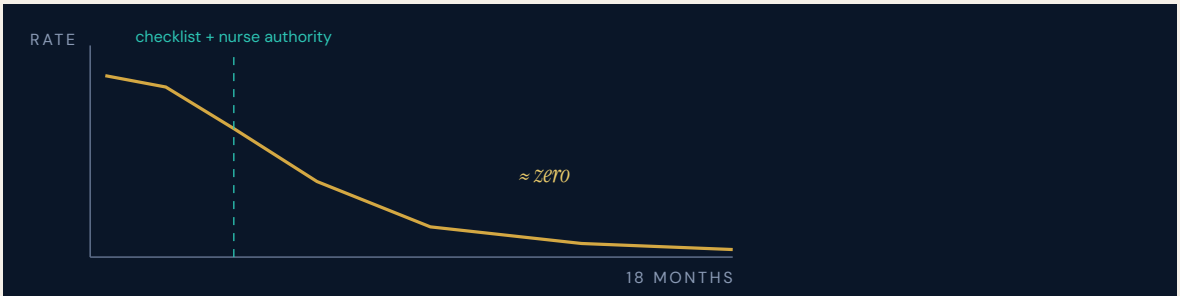
The checklist was the technical half; the nurses' enforcement authority was the cultural half, and it was the load-bearing one. Before Keystone, the path to intervene did not exist; after it, the path was institutional and expected. The pairing converted knowledge everyone already had into behavior that happened every time. The checklist gave the nurse an objective, impersonal basis for the stop — a missed item, not a judgment about the physician — which is what made the authority usable in practice rather than merely on paper.[2]

THE EVIDENCE

Across the Michigan ICUs, the median quarterly CLABSI rate fell to zero, and the state's units outperformed 90% of ICUs nationwide. The program was estimated to have saved roughly 1,500 lives and $175 million within eighteen months. Results were published in the New England Journal of Medicine in 2006, and follow-up showed the effect sustained at ten years. The durability mattered as much as the size of the drop: an improvement that survives a decade is evidence the change was built into the system's structure rather than riding on the initial enthusiasm of a single project.[3]

WHAT TRANSFERRED

Keystone became the clearest evidence in healthcare that a technical intervention without an authority change produces no durable improvement — and vice versa. The model was packaged as the AHRQ CUSP toolkit, adopted in more than forty states, and replicated internationally, establishing the design principle of intervening in matched technical-and-cultural pairs. Packaging the approach as a reusable toolkit was itself part of what transferred — it turned a single project's success into something other institutions could adopt without rediscovering the load-bearing role of the authority change.[4]



BLOODSTREAM-INFECTION RATE ACROSS 103 MICHIGAN ICUS

> *The checklist was the technical intervention. The nurses' authority to enforce it was the cultural intervention. Neither worked without the other.*
>
> PARAPHRASING PRONOVOST & VOHR, SAFE PATIENTS, SMART HOSPITALS, 2010

REFERENCES

1. Pronovost, P. et al. (2006), "An Intervention to Decrease Catheter-Related Bloodstream Infections in the ICU," NEJM 355 — the trial and the near-zero result.
2. Pronovost & Vohr (2010), Safe Patients, Smart Hospitals — the checklist-plus-nurse-authority pairing (paraphrased).
3. Lipitz-Snyderman, A. et al. (2011), BMJ — sustained effect at follow-up.
4. Agency for Healthcare Research and Quality, CUSP toolkit — dissemination across states.
5. Bosk, C. et al. (2009), "Reality check for checklists," The Lancet — the authorization, not the list, as the active ingredient.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Keystone is the clearest evidence in healthcare that a technical intervention without authority intervention produces no durable change, and authority intervention without a technical artifact produces no measurable change. Both are necessary. The empirical record of Keystone is the strongest available argument for designing interventions as **pairs** — and treating the cultural half as engineering, not aspiration.

LENS APPROACH

LENS uses Keystone in LEN 4 and LEN 10 as the canonical worked example of measurement linked to intervention. Studio projects require students to specify both halves of the pair, name the authority that authorizes the cultural half, and identify the measurement signal that will confirm or refute the effect. The course treats "is the cultural change actually authorized?" as falsifiable, not stated.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
HEALTHCARE →	T N →	D1 D2 **D3** D4 D5	3	4.1	3
		Human-System Collaboration			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Pair the technical artifact with the authority change from the outset — a checklist plus the explicit, institution-backed right of a junior member to stop the procedure when a step is missed.
- Make the intervention an obligation rather than a permission, so the corrective action does not depend on an individual's willingness to challenge a senior colleague.
- Anchor the stop to an objective, impersonal trigger (a missed checklist item) so the authority is usable without it reading as a judgment of the more senior operator.

IN OPERATION / AFTER

- Measure the outcome directly (the CLABSI rate) and publish it, so the cultural half can be confirmed as authorized in practice rather than merely declared.
- Track the effect over years, not months, to confirm the change is built into the system's structure rather than riding on a project's initial enthusiasm.
- Package the paired design as a reusable toolkit so other institutions can adopt it without rediscovering the load-bearing role of the authority change.

REFLECTION QUESTIONS

1. Identify an evidence-based protocol in your domain that is **known** to work but is not consistently performed. What is the authority change required to pair with it?

2. Design the measurement signal that would confirm the cultural half of a Keystone-style intervention is actually being authorized — not merely declared.

3. Keystone made the nurse's stop an obligation the institution stood behind, not a personal risk. Identify a corrective action in your domain that currently costs the person who takes it, and design the institutional backing that would remove that cost.

WHO BUILDS THIS learning science & instructional design · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS scenario simulation · competency models · incident analysis · safety dashboards

CASE 20 HEALTHCARE CLINICAL AI 2018 – 2022

TREWS / Sepsis Watch

H Human-System Interface K Knowledge & Institutional Memory G Governance & Trust

DISCLOSURE Institutional overlap: an editor shares an institution (Johns Hopkins) with authors of this work and its deployment sites (five JHU hospitals); no editor was personally involved. Included on the published peer-reviewed evidence (Nature Medicine 2022).

IMPACT **Prospective multi-site evidence of reduced mortality, organ failure, and length of stay when clinicians engaged with ML sepsis alerts in deployed care**

IN BRIEF

The Targeted Real-time Early Warning System (TREWS) is a machine-learning sepsis-detection tool deployed at five Johns Hopkins hospitals; Duke's Sepsis Watch follows the same pattern. The Adams et al. prospective multi-site study (Nature Medicine 2022) reported reduced in-hospital mortality, reduced organ failure, shorter length of stay, and improved antibiotic timeliness associated with deployment — conditional on clinicians acting on alerts within a defined window. The benefit was not the model in isolation. It was the model plus a deliberately engineered alert, workflow, and clinician-confirmation layer that made the alert actionable at the bedside. The honest hedge in the literature is that the evidence is prospective and observational, not RCT; the field notes RCTs are still pending. The case is the positive counter to the Epic Sepsis Model (Case 5): same delegation task, opposite outcome, and the difference is in the engineering of the human–machine boundary and the discipline of the surrounding evidence work.

BACKGROUND

Sepsis is among the most consequential time-dependent diagnoses in hospital medicine: every hour of delayed antibiotics is associated with increased mortality, and the disease is heterogeneous enough that bedside clinicians frequently miss the earliest signal in a patient already being treated for something else. The promise of machine learning has been to surface that earliest signal from the continuously updated EHR trace — vitals, labs, medications — and route it to a clinician who can act in time.[0]

THE INTERVENTION

TREWS, deployed across five Johns Hopkins hospitals, and Sepsis Watch at Duke are the two best-documented instances of this approach. The Adams et al. prospective study of 590,000 patient encounters reported that when clinicians confirmed an alert within three hours, in-hospital mortality, organ failure, and length of stay were lower than for matched controls, and antibiotics were given sooner. The evidence is observational rather than randomized, but it is multi-site, pre-registered, and outcome-grounded.[1]

HOW IT WORKED

The capability the deployment supplied was not "the model." It was the alert designed to fit into a specific clinical workflow, the confirmation step that put a clinician between the model and the action, and the institutional commitment to measure outcomes — not adoption — as the metric of success. The reported benefit collapsed when clinicians did not engage with the alert: the model on its own did nothing. The deliverable was the interface, the role design, and the surrounding evidence loop, not the prediction.[2]

THE EVIDENCE

The honest hedge survives into the literature. The Adams et al. paper, and the broader sepsis-AI field, explicitly note that the outcome inference is conditional on the population, the workflow, and the engagement pattern measured at these sites — and that randomized trials remain pending. The benefit is real on the evidence presented, but it is not a closed proof; it is the strongest available evidence that delegation of early detection to ML can be done as a paired intervention with measurable outcome improvement.[3]

WHAT TRANSFERRED

What TREWS teaches is that the failure pattern of clinical AI (Case 35) is not inevitable. When the model is treated as one component of a deliberately designed human–machine system — with an alert that fits the workflow, a clinician role that retains authority, a deployment that is observed prospectively, and a willingness to report null effects in non-engaged subgroups — the delegation can produce capability rather than alert fatigue. The case is the engineering counter to Watson for Oncology (the model marketed ahead of its capability) and the Epic Sepsis Model (the model deployed ahead of its validation).

> *The benefit is not the model. It is the model used as part of a system clinicians can act on.*
>
> EDITORS' SYNTHESIS OF ADAMS ET AL. (2022) AND HENRY ET AL. (2022).

REFERENCES

1. Adams et al. (2022), "Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis," Nature Medicine 28(7):1455–1460, doi:10.1038/s41591-022-01894-0.
2. Henry et al. (2022), "Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing," Nature Medicine 28(7):1447–1454, doi:10.1038/s41591-022-01895-z.
3. Sendak et al. (2020), "Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study," JMIR Medical Informatics 8(7):e15182 (Sepsis Watch implementation).
4. Wong et al. (2021), "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model," JAMA Internal Medicine 181(8):1065–1070 — the foil case (102).

THE LEARNING ENGINEERING LENS

LE INSIGHT

TREWS is the strongest current evidence that delegating early sepsis detection to a machine-learning system can improve patient outcomes — when the delegation is engineered as a paired intervention. The capability deliverable is the alert design, the clinician role, and the outcome-grounded evidence loop, not the model.

LENS APPROACH

TREWS is the positive Domain 3 / Problem Type 6 case the corpus needed: a documented delegation to AI that worked, with the explanation locatable in the human–machine interface and the governance discipline rather than the model. LENS uses it in Domain 3 (Human-System Collaboration) for the delegation-with-

revocation pattern, in Domain 4 (Test and Evaluation) for outcome-grounded prospective evidence under the judgment-under-inadequate-evidence LEO, and in Domain 5 (Navigating Sociotechnical Constraints) for the workflow-fit discipline. It is the foil drafted directly against the Epic Sepsis Model (Case 5).

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
HEALTHCARE CLINICAL AI	H K G	D1 D2 **D3** D4 D5 Human-System Collaboration	6	3.1	4 · 3

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Design the alert to fit a specific workflow, not the average workflow — including the bedside action it should provoke and the timeline within which it must be confirmed.
- Keep a clinician between the model and the action: model flags, human confirms, the model's authority is to surface, not to decide.
- Instrument the deployment for outcomes (mortality, organ failure, antibiotic timeliness) before the first alert fires — adoption is not the metric of success.

IN OPERATION / AFTER

- Report engagement-stratified outcomes honestly — including the subgroups where the alert was not acted on and the benefit was not observed.
- Treat the prospective/observational design as a constraint to be replaced by RCT evidence when feasible, not as a result that ends the evidence work.
- Carry the model's hedges into the deployment documentation so the next site adopts the model and the discipline that produced it.

REFLECTION QUESTIONS

1. Identify a delegation of detection or screening in your domain that succeeded. What were the components of the human–machine interface that made the model actionable, and what would happen to the outcome metric if those components were removed?

2. Specify the engagement-stratified outcome you would report from a deployment at your site — including the subgroup where the alert was not acted on. What would you need to instrument before the first alert fires?

3. The TREWS evidence is prospective and observational, not RCT. What is the minimum additional evidence you would require before recommending the same model deployment at a new site that differs from the validation sites in population, workflow, or EHR configuration?

WHO BUILDS THIS human factors & interaction design · knowledge management & data engineering · governance, ethics & measurement — plus domain experts and a learning engineer to integrate. TOOLS usability testing · cognitive walkthroughs · knowledge bases · data pipelines · governance frameworks · bias & evidence audits

CASE 26 MEDICAL DEVICES CLINICAL CARE HEALTH EQUITY 1990 – 2020 – 2025

Pulse Oximetry Across Skin Tones

[D] Designed Out [G] Governance & Trust [N] Normalization of Deviance

IMPACT A bedside device validated on a non-representative population systematically under-detected hypoxemia in Black patients for thirty years; the bias persisted because device validation was never demographically stratified

IN BRIEF

Pulse oximetry — the noninvasive bedside measurement of blood oxygen saturation — is one of the most widely used patient-monitoring tools in clinical medicine. Sjoding et al. (NEJM 2020) found that across two large cohorts, Black patients had nearly three times the frequency of **occult hypoxemia** (arterial saturation <88% despite a pulse-ox reading of 92–96%) as White patients. The finding replicated Jubran & Tobin (Chest 1990), published thirty years earlier and overlooked operationally. The bias persisted because device validation was never demographically stratified — the aggregate accuracy number on FDA clearance documentation concealed a group-specific failure mode. The discovery drove FDA review and, in January 2025, a draft guidance recommending that manufacturers evaluate device performance across diverse skin pigmentations during validation. The case is a **failure-to-intervention arc**: the failure sat in the validation architecture for three decades; the intervention is the regulatory change-control on validation, and its measured effect on the disparities outcome is not yet documented.

BACKGROUND

Pulse oximetry replaced repeated arterial blood draws as the bedside standard for monitoring oxygen saturation in the 1980s and 1990s. The device shines two wavelengths of light through tissue and infers saturation from the absorbance ratio. The inference depends, among other things, on the absorbance properties of the intervening tissue — including melanin pigmentation. The clearance documentation reports an aggregate accuracy number against arterial blood gas measurements.[9]

THE INTERVENTION

In 1990 Jubran & Tobin reported, in Chest, that pulse oximeters tended to overestimate true oxygenation in patients with darker skin. The paper was published, cited intermittently, and did not drive a change in validation practice. Thirty years later Sjoding et al. (NEJM 2020) revisited the question in two large modern cohorts and reported that Black patients had nearly three times the frequency of **occult hypoxemia** — true arterial saturation below 88 percent despite a pulse-oximeter reading of 92–96 percent — as White patients. The structural form of the finding was the same as Jubran & Tobin's; the population and the salience

were different, and the COVID-19 pandemic, which made pulse oximetry a household-scale triage tool, made it harder to ignore.[1]

HOW IT WORKED

The bias persisted because device validation was never demographically stratified at clearance. The aggregate accuracy number — clinically acceptable on average — concealed a group-specific failure mode. The capability gap was not in the clinician using the device or in the manufacturer's engineering; it was in the regulatory machinery that approved a measurement device without checking whether its measurement held across the patients it would meet. The Sjoding paper's lasting contribution was not the technical finding alone — Jubran & Tobin had supplied that — but the disconfirmation of the validation architecture.[2]

THE EVIDENCE

The FDA's January 2025 draft guidance recommends that pulse oximeter manufacturers evaluate device performance across diverse skin pigmentations during validation, and that the validation protocol explicitly stratify accuracy by skin tone. The guidance is the corrective-action half of the case: an intervention in the validation architecture, not in the device itself. The measured effect on the downstream disparities outcome — under-treated hypoxemia in patients of color — is not yet documented; it is the continuing work the intervention sets up.[3]

WHAT TRANSFERRED

What the case teaches is that a measurement-device failure can persist for three decades inside an aggregate accuracy number, and that the capability deliverable is a validation architecture that surfaces group-specific failure modes by stratifying outcome metrics by relevant demographic characteristics. Pulse oximetry pairs with eGFR (Case 25) and pain assessment (Case 6) in the race-construct trio. The mechanisms are distinct — eGFR is construct definition; pulse oximetry is validation architecture; Hoffman is clinician-held false belief — and the case-grounded lesson is that the diagnosis of the same surface harm has to attribute the gap to the right layer of the system before a remediation can land.

> *Aggregate accuracy is not group accuracy. A device can be acceptable on average and unsafe for one population.*
>
> EDITORS' SYNTHESIS OF SJODING ET AL. (2020) AND THE FDA 2025 DRAFT GUIDANCE.

REFERENCES

1. Sjoding, Dickson, Iwashyna, Gay, & Valley (2020), "Racial Bias in Pulse Oximetry Measurement," New England Journal of Medicine 383(25):2477–2478, doi:10.1056/NEJMc2029240.
2. Jubran & Tobin (1990), "Reliability of Pulse Oximetry in Titrating Supplemental Oxygen Therapy in Ventilator-Dependent Patients," Chest 97(6):1420–1425 — original finding, published thirty years earlier.
3. FDA (2025), "Pulse Oximeters for Medical Purposes — Non-Clinical and Clinical Performance Testing, Labeling, and Premarket Submission Recommendations: Draft Guidance for Industry and Food and Drug Administration Staff," issued January 7 2025, Docket No. FDA-2023-N-4976; Federal Register notice 2024-31540 — the regulatory corrective-action artifact, language may evolve in final.
4. Fawzy et al. (2022), "Racial and Ethnic Discrepancy in Pulse Oximetry and Delayed Identification of Treatment Eligibility Among Patients With COVID-19," JAMA Internal Medicine — downstream effect during the pandemic.

Pulse oximetry is the validation-architecture instance of the race-construct trio. The bias was published in 1990, persisted for thirty years inside aggregate clearance accuracy, was re-documented in 2020, and reached the regulatory architecture in 2025. The capability deliverable is demographic stratification at validation, not after deployment.

LENS APPROACH

Pulse oximetry is the validation-architecture intervention in the race-construct trio (Cases 25, 26 and 6). LENS uses it in Domain 4 (Test and Evaluation) for the LEO **Gap attribution** — the gap is in the validation architecture, not the device or the clinician — and in Domain 5 (Navigating Sociotechnical Constraints) for the FDA clearance / device oversight regime. Adjacent to Case 35 (radiology AI miscalibration), which is the same diagnosis at a different technological boundary, and to the Epic Sepsis Model (Case 5) for the post-deployment-surveillance pattern.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
MEDICAL DEVICES CLINICAL CARE HEALTH EQUITY	D G N	D1 D2 D3 **D4** D5 Test and Evaluation	5	8.2	4 · 5

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Stratify device-validation outcomes by every demographic characteristic that can change the measurement physics, before clearance, not after deployment.
- Specify the group-specific accuracy that would count as acceptable, separately from the aggregate; do not allow the aggregate to substitute.
- Treat replication of an earlier finding (Jubran & Tobin → Sjoding) as a verification trigger, not a duplication — the same finding in a different population is itself evidence.

IN OPERATION / AFTER

- Audit deployed devices on the population that actually uses them, on a schedule that surfaces drift; aggregate accuracy is not a substitute.
- Tie the regulatory clearance update to a downstream-outcome surveillance plan — under-treated hypoxemia, in this case — so the intervention's effect on the harm can be measured.
- When a published finding sits operationally inert for years, ask whether the publication channel reached the operational community; the structural problem may be in how validation evidence diffuses, not in whether it exists.

REFLECTION QUESTIONS

1. Identify a measurement device in your domain whose validation reports an aggregate accuracy number. Across which demographic characteristics could the measurement physics change? What would a stratified validation protocol look like, and who would have to approve it?

2. The Sjoding finding replicated Jubran & Tobin thirty years later. What is the institutional architecture that should have surfaced the original finding to the regulatory regime? Where did the publication-to-operations channel break, and where does it still break in your domain?

3. The FDA 2025 draft guidance corrects the validation architecture. Specify the downstream outcome (under-treated hypoxemia in patients of color) and the surveillance plan you would tie to the guidance so the intervention's effect on harm is measurable.

WHO BUILDS THIS systems & human-factors engineering · governance, ethics & measurement · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS task analysis · design prototyping · governance frameworks · bias & evidence audits · incident analysis · safety dashboards

Chapter 3

Education, Training & the Learning Workforce — What Fails

When learning systems scale faster than their evidence.

The pilot results were real. The scale-up assumed they would travel on their own.

AN OBSERVATION RECURRING ACROSS THE CHAPTER'S CASES.

CASE 45 EDUCATION AT SCALE 2009–2018

Tennessee Voluntary Pre-K Study

G Governance & Trust D Designed Out

IMPACT Vanderbilt longitudinal RCT of a state-funded universal pre-K program: early gains faded by third grade; sixth-grade outcomes were worse than the control group on several measures

IN BRIEF

Tennessee's Voluntary Pre-K program enrolled some 18,000 four-year-olds a year, and Vanderbilt researchers studied it with a rare large-scale randomized controlled trial — children admitted by lottery against those who applied but didn't enroll. Through kindergarten the pre-K children showed the expected gains in literacy and vocabulary. By third grade the gains had faded, and by sixth grade the pre-K group did somewhat worse than controls on several academic and behavior measures. The unwelcome result was contested, its methods attacked, and the field largely declined to absorb it. The case is in the book not because pre-K is bad — other programs show durable effects — but because the measurement was unusually rigorous and the discipline's capacity to act on an inconvenient finding was tested and largely failed.

BACKGROUND

The Tennessee Voluntary Pre-Kindergarten Program served roughly 18,000 four-year-olds a year. Demand exceeded supply, and that scarcity created an ethical lottery — a fair way to ration scarce seats that doubled as a clean randomizer. It let Vanderbilt's Peabody Research Institute run something rare in education: a randomized controlled trial, following children admitted by lottery against children who applied but did not enroll, the kind of design rarely available in a live policy setting.[o]

WHAT HAPPENED

Through kindergarten the pre-K children showed the expected gains — stronger letter knowledge, vocabulary, early literacy — exactly the early result the program had been funded to produce. By third grade the gains had faded. By sixth grade, the researchers reported, the pre-K children were doing somewhat **worse** than the control group on several state academic measures and on teacher-reported behavior — a reversal that turned an expected success story into an uncomfortable one the longitudinal design had been built to detect.[1]

THE INVESTIGATION

The result contradicted policy consensus and provoked an unusual response: the study was attacked, its methods contested, and the field largely declined to internalize the findings, defending the policy rather than interrogating it. Other rigorous programs — Perry Preschool, Abecedarian — remain durable, so the lesson is not that pre-K fails; it is that a discipline met

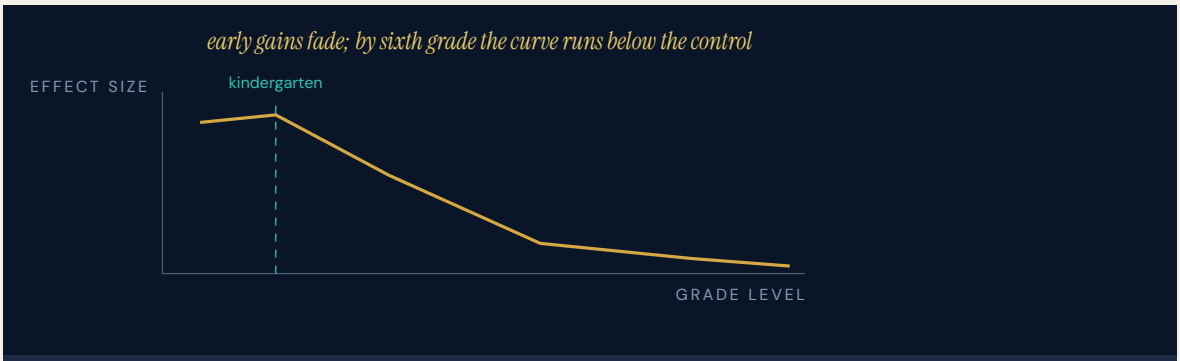
an unwelcome, well-measured answer and mostly looked away, treating the inconvenient evidence as an attack to repel rather than a finding to absorb.[2]

THE CAPABILITY GAP

Tennessee Pre-K is the cleanest case in the dataset for what happens when a field has not engineered its own capacity to update on contrary evidence. The measurement instrument worked — a real RCT in a real policy setting, the rare study whose design could not easily be waved off. The institutional architecture for acting on what it found did not, so a strong instrument fed a discipline with no pathway to receive its answer. A measurement that returns an inconvenient result is only as valuable as the discipline's willingness to absorb it, and here the willingness was the part that was missing.[3]

AFTERMATH & REFORM

The debate continued for years through follow-up studies and counter-analyses, and the episode became a touchstone in the methodology of early-childhood research — argued over more than acted upon.[4] Its place in this book is as a governance-of-evidence case: the capability that needed engineering was not a better study but an implementation-science pathway that could route an unwelcome finding into program redesign rather than rejection, so the cost of the study bought a course correction instead of a controversy.



TENNESSEE PRE-K — MEASUREMENT ARCHITECTURE WORKED; THE FIELD'S RESPONSE DID NOT

> *By sixth grade, children who had attended TN-VPK were doing somewhat worse on academic achievement and discipline measures than children in the control group.*
>
> DURKIN, LIPSEY, FARRAN & WIESEN (2022), VANDERBILT PEABODY RESEARCH INSTITUTE

REFERENCES

1. M. Lipsey, D. Farran & K. Durkin, "Effects of the Tennessee Pre-Kindergarten Program… Through Third Grade," Early Childhood Research Quarterly 45: 155–176 (2018) — the RCT and fade-out.
2. K. Durkin, M. Lipsey, D. Farran & E. Wiesen, "Effects of a Statewide Pre-Kindergarten Program… Through Sixth Grade," Developmental Psychology 58(3): 470–484 (2022) — the sixth-grade reversal (quoted).
3. Responses and counter-analyses to the TN-VPK findings (2018–2022) — the contested reception.
4. National Institute for Early Education Research, State of Preschool yearbooks (2010–2022) — program scale and context.
5. D. Phillips et al., Puzzling It Out: The Current State of Scientific Knowledge on Pre-Kindergarten Effects (2017); J. Heckman on durable early-childhood programs.

LE INSIGHT

Tennessee Pre-K is the cleanest case in the dataset for what happens when a discipline has not engineered its own capacity to update on contrary evidence. The measurement instrument worked. The institutional architecture for acting on what it found did not. Compare to Case 8 (Makary methodology debate) and Case 7 (VA Wait-Time): in each, a measurement that returned an unwelcome answer was contested rather than absorbed.

LENS APPROACH

LENS uses Tennessee Pre-K in LEN 4 as a measurement-architecture case (longitudinal RCT in a real policy context), in LEN 7 to discuss the institutional politics of unwelcome findings, and in LEN 10 as a studio prompt for designing the implementation-science pathway that would absorb such findings into program redesign rather than rejection.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
EDUCATION AT SCALE →	G D →	D1 D2 D3 **D4** D5	5	2.4	4
		Test and Evaluation			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Commit in advance to a decision rule for how an unwelcome but well-measured result will change the program, so the response is designed before the finding arrives.
- Build the longitudinal measurement to detect fade-out and reversal, not just the early gains a program is funded to show.
- Establish an implementation-science pathway that can route a contrary finding into redesign, giving the evidence somewhere to go.

IN OPERATION / AFTER

- Audit how the discipline actually receives inconvenient evidence, treating reflexive method-attacks as a symptom of a missing absorption pathway.
- Sustain follow-up measurement through later grades so a faded or reversed effect cannot be obscured by an early success.
- Protect the independence of the measurement function so a strong instrument is not dismantled for returning an answer the field did not want.

REFLECTION QUESTIONS

1. What measurement instrument in your domain has returned an unwelcome answer, and how did the discipline respond?
2. Design the implementation-science pathway that would absorb a Tennessee-Pre-K-style finding into program redesign rather than rejection.
3. The Tennessee RCT was strong enough that its methods were attacked rather than its conclusion accepted. What decides, in your field, whether a rigorous but inconvenient result is absorbed or contested — and who holds the authority to act on it?

WHO BUILDS THIS governance, ethics & measurement · systems & human-factors engineering — plus domain experts and a learning engineer to integrate. TOOLS governance frameworks · bias & evidence audits · task analysis · design prototyping

CASE 49 GOVERNMENT EDUCATION AT SCALE HIGH-STAKES ASSESSMENT 2020

UK Ofqual A-Level Algorithm — National-Scale Grading Replaced by Algorithm, Withdrawn in Days

D Designed Out K Knowledge & Institutional Memory N Normalization of Deviance

IMPACT Ofqual standardisation algorithm applied to summer 2020 A-level grades following examination cancellation downgraded approximately 39.1% of teacher-estimated grades; results released August 13 2020; algorithm withdrawn August 17 2020 after four days of public protest; Centre Assessment Grades (teacher estimates) substituted; UK House of Commons Education Committee report (2021) documented disproportionate downgrade rates for state-school students

IN BRIEF

With summer 2020 examinations cancelled in response to the COVID-19 pandemic, the UK Office of Qualifications and Examinations Regulation (Ofqual) deployed a statistical standardisation model to produce A-level grades from Centre Assessment Grades (teacher estimates) and Centre-level historical performance. Results were released on August 13,
2020. Approximately 39.1 percent of teacher-estimated grades

were downgraded by the algorithm. State-school students in larger cohorts were downgraded at higher rates than independent- school students in smaller cohorts, because the model relied on Centre-level historical performance more heavily where Centre-level cohorts were larger. After four days of public protest, on August 17, 2020, the algorithm was withdrawn and Centre Assessment Grades were substituted. The Ofqual technical report acknowledges that the standardisation goal was incompatible with population-level fairness given individual-student variance and the dependence of the model on cohort size. The case pairs with Case 86 (Gándara / AERA Open community-college fairness), Case 88 (LiveHint AI bias across foundation models), and Case 48 (Johnson school surveillance).

BACKGROUND

The summer 2020 examination cancellation removed the mechanism by which A-level grades had historically been produced. The Department for Education and Ofqual judged that teacher estimates alone would generate grade inflation incompatible with university admissions and higher-education capacity planning. The decision was to combine teacher Centre Assessment Grades with a statistical standardisation model that drew on Centre-

level historical performance to adjust the distribution of grades each Centre's cohort received. The intent of the standardisation was to preserve year-on-year comparability of the national grade distribution; the seam that surfaced under deployment was that population-level comparability and individual-student fairness are not reconcilable when the standardisation mechanism depends on cohort size.[0]

WHAT HAPPENED

The model's mechanics carried the seam. Where a Centre's cohort for a subject was small, the model relied more heavily on the teacher-submitted Centre Assessment Grade and rank order, because small-cohort historical performance was not informative enough to standardise against. Where a Centre's cohort was large, the model relied more heavily on the Centre's historical performance, because the larger cohort gave the standardisation more purchase. The distributional consequence was structural: independent schools and selective settings with small cohorts received grades close to teacher estimates; state schools and large- cohort comprehensive settings received grades pulled downward toward the Centre's historical distribution. The headline result was that approximately 39.1 percent of teacher-estimated grades were downgraded and that the downgrade rate was higher for state-school students in large cohorts.[1]

THE INVESTIGATION

The four-day withdrawal arc is the governance record. Grades were released on Thursday, August 13, 2020. Public protest — students gathering with signs reading "the algorithm stole my future," extensive press coverage of named individual cases, and rapid political pressure — built across the weekend. On Monday, August 17, 2020, Ofqual and the Department for Education announced that A-level and GCSE grades would be reissued at the teacher-submitted Centre Assessment Grade level. The withdrawal was structural — it affected the entire 2020 cohort — and it was rapid in a way that few national-scale algorithmic deployments have been. The House of Commons Education Committee's 2021 report adjudicated the governance record and named the consultation-with-stakeholders failure as the load-bearing one; the technical report had been internally honest about the cohort-size dependence, and the failure was that the dependence had not been surfaced to affected schools and students in advance of deployment.[2]

THE CAPABILITY GAP

The case pairs with Case 86 (Gándara / community-college predictive equity in AERA Open) at the higher-education scale: both cases turn on the question of whether a standardisation or prediction mechanism that is statistically defensible at the population level can be deployed in a way that is defensible at the individual-student level. Pair with Case 88 (LiveHint AI bias across foundation models) for the bias-surfacing thread in education-deployed algorithms. Pair with Case 48 (Johnson school surveillance) for the algorithmic-administration-in-education parallel at a different scale. The Ofqual case is unusual in the casebook because it is the rare deployment that was withdrawn within days under public pressure; most cases in the corpus document deployments that ran for years before withdrawal or that were never withdrawn.[3]

AFTERMATH & REFORM

The technical report's hedge is binding and load-bearing. Ofqual's own document acknowledges that the standardisation goal — preserving year-on-year comparability of the national distribution — was incompatible with population- level fairness given the variance in individual-student circumstances and the cohort-size dependence of the model. The case teaches the change-control-and-disclosure-as- governance-artifacts pattern: an algorithm that is internally documented as carrying a distributional seam cannot be deployed at population scale without the distributional seam being surfaced to the affected population in advance of deployment. The four-day withdrawal is the governance evidence that the population had not been consulted; the structural argument the case anchors is that the consultation is the governance artifact whose absence the withdrawal made visible.

> *Approximately 39.1 percent of teacher-estimated grades were downgraded; the algorithm was withdrawn within four days under public protest; the technical report was internally honest about the cohort-size dependence and the failure was that the honesty did not travel out of the document.*
>
> EDITORS' SYNTHESIS OF THE OFQUAL TECHNICAL REPORT AND THE HOUSE OF COMMONS EDUCATION COMMITTEE REPORT.

REFERENCES

1. Ofqual, Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: interim report, August 2020.
2. UK House of Commons Education Committee, Getting the grades they've earned: COVID-19: the cancellation of exams and "calculated" grades, HC 254, July 2021.
3. Royal Statistical Society, Submission to the Office for Statistics Regulation on the summer 2020 grading process, 2020 — independent statistical review of the standardisation methodology.
4. Smith, H. (2020), "Algorithmic bias: should students pay the price?" AI & Society 35(4):1077–1078 — early academic commentary on the equity dimensions of the withdrawal.

THE LEARNING ENGINEERING LENS

LE INSIGHT

The Ofqual A-level case is the rare national-scale algorithmic deployment withdrawn within days under public pressure. The technical report was internally honest about the cohort-size dependence and the incompatibility of population-level standardisation with individual-level fairness; the four-day withdrawal is the evidence that the internal honesty did not travel out of the document to the affected population in advance of deployment.

LENS APPROACH

Ofqual A-level 2020 is the change-control-and-disclosure-as- governance-artifacts case at national scale (induced 5.4; LENS D4+D5/PT5; LEO-4 and LEO-5). LENS uses it in Domain 4 (Test and Evaluation) for the consultation-with-affected- stakeholders process as the test surface and in Domain 5 (Navigating Sociotechnical Constraints) for the cohort-size dependence as the distributional seam the deployment carried. Pair with Case 86 (Gándara community-college predictive equity), Case 88 (LiveHint AI bias across foundation models), and Case 48 (Johnson school surveillance). The technical report's acknowledgement of incompatibility is the load-bearing hedge.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
GOVERNMENT EDUCATION AT SCALE HIGH-STAKES ASSESSMENT	D K N	D1 D2 D3 **D4** D5 Test and Evaluation + D5	5	5.4	4 · 5

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Surface the distributional seam an internally documented standardisation mechanism carries to the affected population in advance of deployment; the Ofqual technical report was internally honest about the cohort-size dependence, and the governance failure was that the honesty did not travel out of the document.
- Treat the consultation-with-affected-stakeholders process as the change-control artifact a national-scale algorithmic deployment requires; the four-day withdrawal under public protest is the evidence that the consultation had not occurred.
- Pre-specify the individual-student fairness criterion against which a standardisation mechanism will be evaluated, and refuse deployment when the criterion is incompatible with the standardisation goal.

IN OPERATION / AFTER

- Carry the technical report's hedge — "standardisation incompatible with population-level fairness given individual-student variance" — into print without softening; the case's pedagogical value depends on the internal documentation of the seam being visible alongside the public withdrawal.
- Pair in syllabi with Case 86 (Gándara) so the population-level-versus-individual-level fairness tension is taught at both the secondary-to-higher-education transition scale and the community-college transition scale.
- Use the case as the rare example of an algorithmic deployment withdrawn at national scale within days; the four-day withdrawal arc is the curricular target for governance-response speed under public pressure.

REFLECTION QUESTIONS

1. Identify a standardisation or prediction mechanism in your domain whose internal documentation flags a distributional seam. What is the consultation process that would surface the seam to the affected population in advance of deployment, and what would deployment without consultation look like?

2. Specify the individual-level fairness criterion against which a population-level standardisation in your setting would be evaluated. Is the criterion compatible with the standardisation goal, and if not, which governs?

3. The Ofqual case is unusual for the speed of withdrawal — four days. Pick a deployment in your domain whose distributional seam has not yet surfaced publicly, and ask: what would have to be true for a four-day withdrawal arc to be possible, and what would have to be true for the seam to have been surfaced in advance instead?

WHO BUILDS THIS systems & human-factors engineering · knowledge management & data engineering · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS task analysis · design prototyping · knowledge bases · data pipelines · incident analysis · safety dashboards

CASE 53 EDUCATION AT SCALE 2014

inBloom

G Governance & Trust

IMPACT **$100M initiative collapsed in ~14 months; all 9 announced partner states pulled out or disavowed participation; data infrastructure for education set back years**

IN BRIEF

inBloom was a $100-million, Gates- and Carnegie-funded shared data infrastructure for U.S. student records — and the technology worked: no bug, no breach, no performance failure. What killed it in about fourteen months was everything around the technology. It launched without consent frameworks, community engagement, transparency about data use, or a way for parents to participate in decisions about their children's data. Parent-privacy groups organized opposition state by state, and nine states withdrew. Analysts read inBloom as the failure of technocratic reform — the assumption that technically sound infrastructure generates its own legitimacy. It is the purest governance failure in the dataset, and the book's clearest argument that in education at scale, stakeholder trust and governance are not optional features but load-bearing structure.

BACKGROUND

inBloom was an ambitious shared data infrastructure for U.S. K-12 student records, backed by roughly $100 million from the Gates Foundation and the Carnegie Corporation of New York, hosted on commercial cloud, and built by enterprise engineers. The premise was that a common store would spare districts from rebuilding the same plumbing and let applications interoperate across systems that had never spoken to one another. Technically it was sound — no bug, no breach, no performance failure ever undid it, and that very soundness is what makes the case instructive.[0]

WHAT HAPPENED

What undid it was everything around the technology. inBloom launched without adequate consent frameworks, without meaningful community engagement on data governance, without transparency about what was collected and why, and without any way for parents to participate in decisions about their children's data. Each omission read, to a worried parent, as a decision made about their child without them in the room. Parent groups organized opposition state by state, and nine states — among them Louisiana and Illinois — withdrew as the political cost of staying overtook any promised efficiency. New York, the last and largest partner, was barred from sharing student data with inBloom by a provision in the state budget enacted at the end of March 2014, and on 21 April 2014 inBloom announced it would wind down. Within about fourteen months the $100-million initiative had collapsed.[1]

THE INVESTIGATION

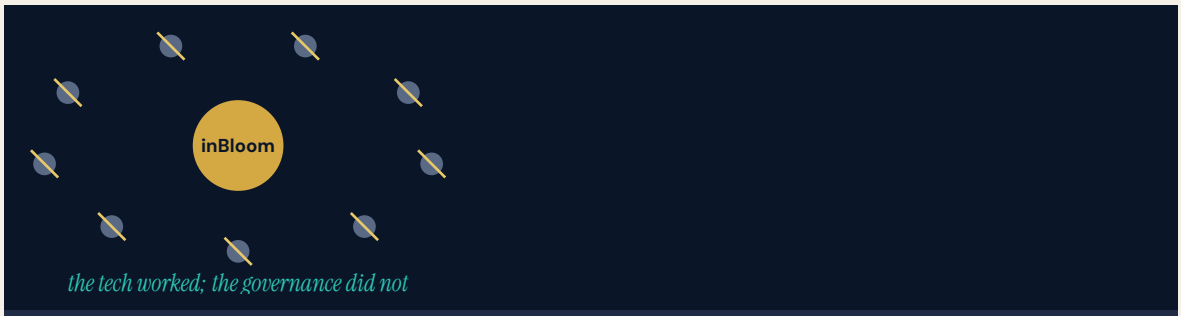
Analysts at Data & Society read inBloom as the failure of technocratic education reform: the assumption that technically sound infrastructure generates its own legitimacy proved catastrophically wrong. Legitimacy, on this reading, is earned from the stakeholders a system acts upon, not conferred by the quality of its engineering. The technology was never the problem; the governance was — the consent, transparency, and trust that had been treated as add-ons rather than as the foundation the whole effort needed before a single record moved.[2]

THE CAPABILITY GAP

inBloom is the purest governance failure in this dataset: nothing technical was wrong, and everything sociotechnical was. The missing capability was the design of stakeholder trust — consent, accountability, and a voice for the families whose data was at stake — treated as a precondition for deployment rather than a feature to add later. Once opposition formed, no patch could retrofit the trust that should have been built in from the start, because trust withheld at launch cannot be engineered back in under fire. In education at scale, those are load-bearing elements, not optional ones.[3]

AFTERMATH & REFORM

inBloom's collapse set back shared education-data infrastructure for years and became a standard cautionary tale; it also helped drive a wave of state student-data-privacy laws that codified, after the fact, the consent and transparency the project had skipped.[4] The lesson the book takes from it is that ethics-as-design-constraint is not ideology but engineering — and inBloom is the $100-million empirical test of what happens when you skip it, paid not in downtime but in a dead initiative and a chilled field.



$100M OF WORKING TECHNOLOGY; NINE STATES WITHDREW

> *The technology was not the problem. The governance was the problem.*
>
> PARAPHRASING THE ANALYSIS IN BULGER, MCCORMICK & PITCAN, DATA & SOCIETY, 2017

REFERENCES

1. M. Bulger, P. McCormick & M. Pitcan, The Legacy of inBloom, Data & Society Research Institute (2017) — inBloom as a failure of technocratic reform.
2. Education Week and Hechinger Report coverage of the state withdrawals (2013–2014) — nine states exiting.
3. Bulger et al. (2017) — governance, not technology, as the cause; "the technology was not the problem."
4. N. Selwyn, Distrusting Educational Technology (2014); d. boyd & K. Crawford, "Critical Questions for Big Data" (2012).
5. Parent Coalition for Student Privacy archives and the wave of state student-data-privacy legislation that followed inBloom.

THE LEARNING ENGINEERING LENS

LE INSIGHT

inBloom is the purest governance failure in this dataset. Nothing technical was wrong. Everything sociotechnical was. The case is the strongest argument in the book for treating ethics-as-design- constraint not as ideology but as engineering: a $100M empirical test of what happens when you do not.

LENS APPROACH

LENS uses inBloom in LEN 7 as the canonical governance failure and in LEN 1 as a stakeholder-analysis case. The case anchors the LENS threading commitment that equity and accountability are design commitments, not modules. Studio projects (LEN 10) require students to produce a stakeholder-trust deliverable as a precondition for deployment.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
EDUCATION AT SCALE →	G →	D1 D2 D3 D4 **D5** Navigating Sociotechnical Constraints	4	5.1	5

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Treat consent, transparency, and a parent-facing voice as load-bearing requirements gathered before the data store is built, not features bolted on after launch.
- Engage the families and districts whose data is at stake as design stakeholders from the outset, so the governance questions surface in requirements rather than in opposition.
- Make legitimacy an explicit deliverable: document who must agree, on what terms, before any student record moves into shared infrastructure.

IN OPERATION / AFTER

- Audit live deployments for the gap between technical soundness and stakeholder trust, since a clean system can still be losing the political ground it stands on.
- Monitor state-by-state consent and withdrawal signals as a leading indicator, treating organized parent opposition as data about a governance defect, not noise.
- Sustain a standing transparency channel so families can see what is collected and why throughout operation, not only at adoption.

REFLECTION QUESTIONS

1. What is the equivalent unbuilt governance infrastructure in your domain? What would the $100M empirical test of its absence look like?
2. Design the stakeholder-trust deliverable that a future inBloom-equivalent should have to produce before deployment.
3. inBloom's engineers were excellent and its technology never failed, yet the project collapsed. Where in your own domain is technical soundness being mistaken for legitimacy — and who would have to consent before a system you build could claim it?

WHO BUILDS THIS governance, ethics & measurement — plus domain experts and a learning engineer to integrate.
TOOLS governance frameworks · bias & evidence audits

CASE 60 K-12 EDUCATION TEACHER EVALUATION ALGORITHMIC ACCOUNTABILITY 2011–2017

Houston EVAAS — Value-Added Teacher Evaluation on Trial

G Governance & Trust

IMPACT Houston ISD used the proprietary SAS EVAAS value-added model to attribute student test-score growth to individual teachers and made termination decisions on the scores during the 2011–15 school years; SAS held the model as a trade secret, so teachers could not inspect, replicate, or contest their scores; in May 2017 the federal district court in Hous. Fed'n of Teachers v. Hous. Indep. Sch. Dist., 251 F. Supp. 3d 1168 (S.D. Tex.), denied summary judgment, holding teachers had a protectable property interest and that unverifiable scores could violate Fourteenth Amendment procedural due process; HISD settled in October 2017, agreeing not to use unverifiable value-added scores as a basis for termination and paying $237,000 in fees

IN BRIEF

From 2011 the Houston Independent School District used the SAS Education Value-Added Assessment System (EVAAS), a proprietary statistical model, to attribute student test-score growth to individual teachers, and made termination decisions on the resulting scores. SAS held the model as a trade secret, so teachers could not inspect, replicate, or contest the number that ended their careers. In Houston Federation of Teachers v. Houston ISD, the federal district court denied summary judgment in May 2017, holding that teachers had a protectable property interest and that unverifiable scores could violate Fourteenth Amendment procedural due process — a "black box" evaluation denies any meaningful opportunity to contest. HISD settled in October 2017, agreeing to stop using unverifiable value-added scores for termination and paying $237,000 in fees. The research record — year-to-year score instability, the ASA 2014 and AERA 2015 statements, and documented subgroup-composition bias — makes this the casebook's cleanest adjudicated gap-attribution failure.

BACKGROUND

Beginning in 2011, the Houston Independent School District tied high-stakes teacher-appraisal decisions — including bonuses and terminations — to the SAS Education Value-Added Assessment System (EVAAS), a proprietary value-added model (VAM) that estimates an individual teacher's contribution to student growth on standardized tests. The operational theory was direct: student test-score growth, statistically adjusted, isolates the teacher's effect, and the isolated effect is decision-grade evidence for employment action.

The statistical profession disagreed on both counts. The American Statistical Association's 2014 statement cautioned that VAMs measure correlation rather than causation, are sensitive to model and data choices, and that most VAM studies find teachers account for about 1% to 14% of the variability in test scores — with the majority of improvement opportunities in system-level conditions. The AERA's 2015 statement enumerated eight technical requirements for valid VAM use and cautioned against giving VAM dispositive weight in high-stakes evaluation. HISD's deployment preceded both statements and continued after them.[0]

WHAT HAPPENED

Teachers rated ineffective by EVAAS could be, and were, exited on the strength of the score, but no teacher could verify it. SAS classified the model's source code and implementation as a trade secret and refused to release them to the district or its employees; the score arrived as a number without a derivation. The independent research record gave teachers concrete reasons to want one. Amrein-Beardsley and Collins's 2012 study of EVAAS in Houston documented teachers reporting that their instruction changed little from year to year while their EVAAS scores swung widely — roughly half of surveyed teachers reported consistent classifications across years, about what a coin flip would produce — and principals reported pressure to align their observational ratings with the EVAAS number. A score that is unstable year-to-year, generated by a model no one outside the vendor can examine, was the sole or decisive input to termination decisions.[1]

THE INVESTIGATION

The Houston Federation of Teachers and seven teachers sued in 2014. In May 2017, the federal district court (S.D. Tex., Magistrate Judge Stephen Wm. Smith) denied HISD summary judgment on the procedural due process claim, holding that teachers had a constitutionally protectable property interest in continued employment and that the EVAAS scores' unverifiability could deny them due process: "HISD teachers have no meaningful way to ensure correct calculation of their EVAAS scores, and as a result are unfairly subject to mistaken deprivation of constitutionally protected property interests in their jobs." The opinion noted that the score "might be erroneously calculated for any number of reasons, ranging from data entry mistakes to glitches in the computer code itself," and that algorithms are human creations, subject to error like any other human endeavor. In October 2017 HISD settled: the district agreed not to use unverifiable value-added scores as a basis for termination so long as they remain unverifiable, created a joint instructional consultation panel on the appraisal process, and paid $237,000 in legal fees.[2]

THE CAPABILITY GAP

The capability gap sits at the attribution step. Student test-score growth is a system outcome — shaped by curriculum, student assignment patterns, school effects, peer composition, and prior schooling — and EVAAS attributed it to an individual operator, the teacher, through a statistical coupling that was never validated for the decision it was driving. The ASA's 1%-to-14% figure is the quantitative form of the gap: the model assigned to the individual a residual dominated by the system. The evidence then failed the decision-grade test so completely that a federal court said so — not that the model was inaccurate, which no one could establish either way, but that a score no affected party can inspect, replicate, or contest cannot support a consequential deprivation. The failure is adjudicated gap misattribution: the institution located a system-level shortfall in the individual because

the individual was the addressable unit, and defended the attribution with secrecy rather than validation.[3]

AFTERMATH & REFORM

The bias record is documented, not asserted, and its hedges are load-bearing. Amrein-Beardsley and Geiger's 2020 analysis of more than 1,700 HISD teachers' EVAAS results found that teachers in schools serving larger populations of English-language learners, free-or-reduced-lunch students, and special-education students received systematically lower value-added scores — a school-composition correlation, which is evidence of bias against teachers serving those populations but not a teacher-level causal decomposition. Concerns about teachers of highly mobile student populations appear in the Houston survey research and practitioner record but are less precisely quantified in the published literature. The settlement did not invalidate value-added modeling; it established that a proprietary, unverifiable score cannot be the basis for termination, and the due-process holding has since anchored the legal literature on algorithmic accountability in public employment. Houston reads, in retrospect, as the leading edge of a broader retreat: after the federal Every Student Succeeds Act (2015) removed the test-based-evaluation mandate that had driven adoption, high-stakes value-added teacher evaluation was rolled back across most of the roughly forty states that had embraced it — the national policy following the direction the ASA's 2014 warning had pointed. Pair with Case 51 (Atlanta Public Schools) for the adjacent measurement-architecture failure and Case 50 (Wisconsin DEWS) for the education-prediction equity thread.[4]

> *When a public agency adopts a policy of making high stakes employment decisions based on secret algorithms incompatible with minimum due process, the proper remedy is to overturn the policy.*
>
> HOUS. FED'N OF TEACHERS V. HOUS. INDEP. SCH. DIST., 251 F. SUPP. 3D 1168 (S.D. TEX. 2017)

REFERENCES

1. Hous. Fed'n of Teachers v. Hous. Indep. Sch. Dist., 251 F. Supp. 3d 1168 (S.D. Tex. 2017) — memorandum opinion denying summary judgment on the procedural due process claim.
2. Amrein-Beardsley, A., & Collins, C. (2012), "The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and Unintended Consequences," Education Policy Analysis Archives 20(12), doi:10.14507/epaa.v20n12.2012.
3. American Statistical Association (2014), ASA Statement on Using Value-Added Models for Educational Assessment, April 8, 2014.
4. American Educational Research Association Council (2015), "AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs," Educational Researcher 44(8):448–452, doi:10.3102/0013189X15618385.
5. Amrein-Beardsley, A., & Geiger, T. (2020), "Methodological Concerns About the Education Value-Added Assessment System (EVAAS): Validity, Reliability, and Bias," SAGE Open 10(2), doi:10.1177/2158244020922224.
6. Education Week (October 2017), "Houston District Settles Lawsuit With Teachers' Union Over Value-Added Scores"; Houston Public Media (October 10, 2017), "Federal Lawsuit Settled Between Houston's Teacher Union and HISD" — settlement terms and the $237,000 fee figure.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Houston EVAAS is the casebook's cleanest adjudicated gap-attribution failure: a system outcome — student growth, shaped by curriculum, assignment patterns, and school effects — was attributed to an individual operator through a coupling never validated for the termination decisions it drove, and the evidence failed the decision-grade test so completely that a federal court said so. The fairness finding — lower scores for

teachers serving larger ELL, free-or-reduced-lunch, and special-education populations — is documented, not asserted, and travels as a school-composition correlation.

LENS APPROACH

Houston EVAAS is the unvalidated-attribution-under-secrecy case (induced 2.1; LENS D4/PT4; LEO-4). LENS uses it in Domain 4 (Test and Evaluation) as the anchor for the decision-grade-evidence standard: the score driving the consequential decision was unverifiable by design, unstable year-to-year, and assigned to the individual a variance residual the ASA's own statement located overwhelmingly in the system. The court's due-process holding gives the standard its adjudicated form — evidence no affected party can inspect, replicate, or contest cannot support a consequential deprivation. Pair with Case 51 (Atlanta Public Schools) for the measurement-architecture thread and Case 50 (Wisconsin DEWS) for the education-prediction equity thread.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
K-12 EDUCATION TEACHER EVALUATION ALGORITHMIC ACCOUNTABILITY	G	D1 D2 D3 **D4** D5 Test and Evaluation	4	2.1	4

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Validate the attribution before the deployment: a model that assigns a system outcome (student growth) to an individual operator (the teacher) must be validated for the specific consequential decision it will drive, not just for aggregate statistical fit — and the ASA's system-dominated variance decomposition is the null hypothesis the validation must beat.
- Treat contestability as a design requirement for any consequential score: if the affected party cannot inspect, replicate, or meaningfully contest the number, the evidence is not decision-grade regardless of the model's internal quality, and a trade-secret claim does not waive the requirement.
- Take the professional bodies' statements as binding inputs: the ASA 2014 and AERA 2015 statements enumerated the technical conditions for VAM use in high-stakes evaluation, and a deployment that cannot meet them should not carry dispositive weight in employment action.

IN OPERATION / AFTER

- Use the adjudicated record as the teaching text: the court's holding — that unverifiable algorithmic scores deny a meaningful opportunity to contest — is the clearest available statement of the decision-grade-evidence standard applied to an operating deployment.
- Audit the equity record at the school-composition level and preserve its hedges: the documented finding is that teachers serving larger ELL, free-or-reduced-lunch, and special-education populations scored systematically lower, and the finding travels as a composition correlation, not a causal decomposition.
- Carry the settlement's structural remedy forward: the joint instructional consultation panel — affected-party representation inside the appraisal-design process — is the governance form that the original deployment lacked.

REFLECTION QUESTIONS

1. Identify a metric in your domain that attributes a system-shaped outcome to an individual operator for a consequential decision. What validation evidence exists for the attribution step specifically — not for the model's aggregate fit — and what fraction of the outcome's variance does the individual plausibly control?

2. Specify the contestability requirement for a consequential score in your setting: what would an affected party need — data, code, derivation, an appeal surface — to meaningfully verify or contest the number, and does any trade-secret or proprietary claim currently block it?

3. The Houston court found the due-process violation without resolving whether EVAAS was accurate, because no one could establish that either way. Pick a proprietary evaluation system in your domain and ask: if its accuracy can be neither confirmed nor refuted by the people it judges, what decisions is it currently supporting, and what is the institution's documented answer for why that is acceptable?

WHO BUILDS THIS governance, ethics & measurement — plus domain experts and a learning engineer to integrate.
TOOLS governance frameworks · bias & evidence audits

CASE 61 K-12 EDUCATION LITERACY INSTRUCTION RESEARCH-PRACTICE GAP 1997–2023

Sold a Story — The Science of Reading and the Decades-Long Research-Practice Gap

K Knowledge & Institutional Memory T Training Gap

EVIDENCE TIER journalism-tier — source confidence flagged; future validation ongoing.

IMPACT National Reading Panel (2000) and decades of peer-reviewed cognitive science established systematic phonics as essential to early reading instruction; classrooms across the United States instead deployed balanced-literacy approaches built on three-cueing, institutionalized in popular curricula (Fountas & Pinnell, Calkins Units of Study, Reading Recovery); Emily Hanford's APM Reports investigations — "Hard Words" (2018) and "Sold a Story" (October 2022) — traced the divergence to teacher preparation and curriculum publishing; by late 2024 roughly 40 states plus DC had passed evidence-based reading laws, Columbia Teachers College dissolved the Reading and Writing Project (September 2023), and New York City mandated replacement curricula (May 2023); student-outcome effects of the legislative wave are still emerging

IN BRIEF

The peer-reviewed evidence base on early reading — decades of cognitive science consolidated by the congressionally mandated National Reading Panel report in 2000, and restated for a practitioner audience by Castles, Rastle, and Nation in 2018 — established systematic, explicit phonics instruction as an essential component of teaching children to read. Across the same decades, a large share of American classrooms deployed balanced-literacy approaches built on three-cueing (meaning, structure, visual — MSV), which directs early readers toward pictures, context, and syntax to identify words, institutionalized in the era's most widely adopted curricula and interventions: Fountas & Pinnell, Lucy Calkins's Units of Study, and Reading Recovery. Emily Hanford's APM Reports investigations — "Hard Words" (2018) and the podcast series "Sold a Story" (launched October 20, 2022) — assembled the reconstruction of how the evidence failed to couple to the operator layer: teacher-preparation programs that did not teach the research, curricula that institutionalized contradicted methods, and an institution that had forgotten that the reading wars had been substantially settled before. The reporting was followed by a wave of state legislation — roughly 40 states plus the District of Columbia with evidence-based reading laws or policies by late 2024 — and by curriculum withdrawals and revisions. The evidence-tier flag is binding: the NRP findings are peer-reviewed, but the causal narrative of why practice diverged is journalistic synthesis, and the student-outcome effects of the legislative wave are still emerging.

BACKGROUND

The evidence base predates the failure by decades. Jeanne Chall's 1967 synthesis had already concluded that code-emphasis instruction outperformed meaning-emphasis approaches for beginning readers, and the accumulating cognitive science of the 1980s and 1990s converged on the same architecture: skilled reading runs through letter-sound decoding, and beginners need systematic, explicit instruction in the code. The National Reading Panel, convened by Congress and reporting in 2000, reviewed the experimental literature and found systematic phonics instruction significantly more effective than unsystematic or no-phonics alternatives for early readers. Castles, Rastle, and Nation's 2018 review — titled, pointedly, "Ending the Reading Wars" — restated the consensus for a practitioner and policymaker audience. The scientific question of how children learn to read words was, by the field's own account, not open during the years the divergent practice scaled.[0]

WHAT HAPPENED

What scaled in classrooms instead was balanced literacy built on three-cueing: the instructional theory that early readers identify words by coordinating meaning, structure, and visual cues — pictures, context, syntax, first letters — rather than by decoding letter by letter. The approach descended from Marie Clay's Reading Recovery and Ken Goodman's "psycholinguistic guessing game," and it was institutionalized in the most commercially successful curricula of the era: Fountas & Pinnell's leveled-literacy systems and Lucy Calkins's Units of Study, the latter used before the pandemic by roughly half of New York City elementary schools that reported a curriculum. Hanford's reporting — "Hard Words" in 2018, then the "Sold a Story" podcast series launched October 20, 2022 — documented the cueing practice in classrooms, traced its lineage, and put to its publishers and authors the question the evidence base had already answered. The series' reconstruction is journalistic synthesis, and the case carries it as such; but the classroom practice it documented was not in serious dispute.[1]

THE INVESTIGATION

The analytical spine of the case is the uncoupled operator layer. The National Reading Panel report was a federal consensus document; it did not, by itself, change what teacher-preparation programs taught. Hanford's reporting, and the National Council on Teacher Quality's program reviews alongside it, found that large numbers of preparation programs did not teach the settled science — graduating teachers who had never encountered the evidence their profession had commissioned. The curricula then institutionalized the contradicted method at the point of use, so that a teacher's daily materials argued against the research even where the teacher had met it. This is implementation science's core question — does the practice hold when it moves into real settings? — inverted: the practice never moved at all. And it is a knowledge-loss failure as much as a training-gap failure: the reading wars had been substantially settled before, in 1967 and again in 2000, and the institution forgot its own evidence each time the settlement failed to reach the operator pipeline.[2]

THE CAPABILITY GAP

The reporting was followed by institutional movement at a pace the evidence alone had not produced in two decades. By Education Week's tracking, roughly 40 states plus the District of Columbia had passed laws or adopted policies on evidence-based reading instruction by late 2024 — addressing curricula, teacher preparation, coaching, and intervention — with

about 15 states acting in 2024 alone. New York City announced in May 2023 that its elementary schools would be required to adopt one of three approved reading curricula, with the chancellor stating publicly that the balanced-literacy approach "has not worked." Columbia Teachers College announced in September 2023 that it would dissolve the Teachers College Reading and Writing Project, the consultancy Calkins founded and directed. Calkins had removed cueing from the 2022 revision of Units of Study and continued her work through a successor organization; Heinemann and Fountas & Pinnell disputed that their materials constituted three-cueing while adding decodable texts to subsequent editions. None of the authors or publishers conceded the reporting's central account.[3]

AFTERMATH & REFORM

The hedges the case carries are load-bearing. The evidence-tier flag is binding: the NRP findings and the reading science are peer-reviewed, but the causal narrative of why practice diverged — the publishing economics, the teacher-preparation gaps, the role of specific authors — is journalistic synthesis, attributed to the reporting rather than to an adjudicated or systematic record. The measured student-outcome effects of the legislative wave are still emerging: Mississippi's NAEP fourth-grade reading gains after its 2013 Literacy-Based Promotion Act are suggestive and widely cited, but the attribution is contested — the law bundled curriculum, coaching, retention, and screening, and the contribution of each is not isolated. Whether the 40-state wave produces population-scale reading gains is an open empirical question, and the case must be taught with that question open. What the case establishes without hedge is the structural finding: an evidence base can be settled, federally consolidated, and publicly available for decades while the operator layer — teacher preparation and daily curriculum — carries the contradicted practice at national scale.

> *The scientific question of how children learn to read words was, by the field's own account, settled while the divergent practice scaled; the evidence base failed to reach teacher preparation and daily curriculum for decades, and the causal narrative of why is journalistic synthesis — source confidence flagged, future validation ongoing.*
>
> EDITORS' SYNTHESIS OF HANFORD (APM REPORTS, 2018, 2022), THE NATIONAL READING PANEL REPORT (2000), AND CASTLES, RASTLE & NATION (2018).

REFERENCES

1. Hanford, E. (2022 – 2024), "Sold a Story: How Teaching Kids to Read Went So Wrong," APM Reports, podcast series launched October 20, 2022.
2. Hanford, E. (2018), "Hard Words: Why Aren't Our Kids Being Taught to Read?" APM Reports, September 10, 2018.
3. National Reading Panel (2000), Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction, National Institute of Child Health and Human Development, NIH Pub. No. 00-4769.
4. Castles, A., Rastle, K., & Nation, K. (2018), "Ending the Reading Wars: Reading Acquisition From Novice to Expert," Psychological Science in the Public Interest 19(1):5 – 51.
5. Schwartz, S. (updated through 2024), "Which States Have Passed 'Science of Reading' Laws? What's in Them?" Education Week — legislation tracker; 40 states plus DC by late 2024.
6. Seidenberg, M. (2017), Language at the Speed of Sight: How We Read, Why So Many Can't, and What Can Be Done About It, Basic Books.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Sold a Story is the load-bearing case for an evidence base that never moved into deployment: decades of peer-reviewed reading science, federally consolidated in 2000, coexisted at national scale with class-

room practice built on the contradicted method, because the operator layer — teacher preparation and curriculum — was never coupled to the evidence. The knowledge-loss reading travels with the training-gap reading: the settlement had been reached before and forgotten, and it took journalism, not the field's own feedback channels, to surface the gap.

LENS APPROACH

Sold a Story is the research-practice-coupling case at national scale (induced 6.4; LENS D2/PT2; LEO-2). LENS uses it in Domain 2 (Iterative Development) for the discipline of treating evidence-to-deployment coupling as the deliverable — implementation science's core question inverted, because the practice never moved at all — and for the operator-pipeline-as-deployment-surface framing. Pair with Case 50 (Wisconsin DEWS) and Case 51 (Atlanta Public Schools) for the education-at-scale evidence thread. The evidence-tier flag is binding: the reading science is peer-reviewed, the causal narrative is journalistic synthesis, and the outcome effects of the legislative wave are still emerging.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
K-12 EDUCATION LITERACY INSTRUCTION RESEARCH-PRACTICE GAP	K T	D1 **D2** D3 D4 D5 Iterative Development	2	6.4	2

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Treat the operator pipeline as part of the deployment surface; an evidence consensus that does not reach teacher preparation and daily curriculum has not been deployed, and the coupling — not the consensus — is the deliverable to verify.
- Audit the installed base against the evidence base on a cycle; the reading wars had been substantially settled twice before the divergent practice scaled, and the institution had no mechanism that compared what its classrooms ran against what its research had concluded.
- Instrument the point of use, not the policy layer; the contradicted method survived in curricula and classroom routines long after the consensus documents shipped, and only observation at the operator layer — what the materials actually direct teachers to do — surfaced the gap.

IN OPERATION / AFTER

- Carry the evidence-tier discipline into teaching; the peer-reviewed reading science and the journalistic causal narrative are different confidence classes, and the case's value depends on students learning to hold both without collapsing either.
- Track the legislative wave as an open experiment; roughly 40 states changed law faster than the evidence changed practice in twenty years, and the population-scale outcome data now accumulating is the test of whether mandate-led coupling works.
- Use the case as the anchor for research-practice coupling in any domain with an operator pipeline; the structural finding — settled evidence, uncoupled preparation layer, institutionalized contradicted practice — generalizes beyond literacy.

REFLECTION QUESTIONS

1. Identify an evidence consensus in your domain that has been settled for a decade or more. Map the operator pipeline — the preparation programs, the daily materials, the point-of-use routines — and ask at which layer the consensus stops. What would the Hanford-style audit of your domain's classrooms find?

2. The reading wars were substantially settled in 1967 and again in 2000, and the settlement was lost each time. Specify the institutional memory mechanism your domain would need for a settled question to stay settled at the operator layer — and name who currently owns it, if anyone.

3. The corrective wave arrived through journalism and legislation rather than through the field's own channels, and its student-outcome effects are still emerging. What is the disciplined position between "the mandate fixed it" and "nothing is known" — and what evidence, on what timeline, would move you off it?

WHO BUILDS THIS knowledge management & data engineering · learning science & instructional design — plus domain experts and a learning engineer to integrate. TOOLS knowledge bases · data pipelines · scenario simulation · competency models

Chapter 4

Education, Training & the Learning Workforce — What Works — and the Frontier

When instruction, analytics, and workforce design run the full evidence loop.

The effect survived because somebody kept measuring after the launch.

AN OBSERVATION RECURRING ACROSS THE CHAPTER'S CASES.

CASE 68 WORKFORCE DEV STEM TRAINING EQUITY 2017 – 2023 (SIX CYCLES)

CIRCUIT — A Scalable, Equity-Centered Research Workforce Model

T Training Gap K Knowledge & Institutional Memory

DISCLOSURE Authorship: an editor of this volume is the senior author of the underlying study. Included on the published peer-reviewed evidence (ASEE 2023); editorial framing keeps critical distance from the program's self-presentation, and the open question about external/comparative evaluation is preserved in the text.

IMPACT **An eight-pillar cohort model that in 2022 supported over 100 undergraduate, graduate, and ROTC students from 'trailblazing' backgrounds (first-generation, low-income, limited prior research access); peer-reviewed program description with longitudinal student outcomes across six cycles**

IN BRIEF

CIRCUIT is a research workforce-development program at Johns Hopkins APL built on eight explicit pillars — holistic recruiting, mission-driven research, targeted technical training, leadership development, high-resolution assessment, diverse mentorship, university partnerships, and career empowerment. Cervantes, Floryanzia, Sharp, Gray-Roncal, and Johnson ("Empowering Trailblazers toward Scalable, Systematized, Research-Based Workforce Development," ASEE Annual Conference 2023) presents the model and reports longitudinal student outcomes gathered over six program cycles. In 2022 the program supported over 100 undergraduate, graduate, and ROTC students, positioning CIRCUIT as a replicable model for STEM recruitment and retention of underrepresented students. The strongest honest framing — preserved in the case — is that this is a self-authored multi-cycle program evaluation at a single program; an external comparative evaluation would strengthen the causal claim, and the case says so rather than overstate. The COI render under the title (editor is the senior author) is binding. The case is the paired peer-reviewed companion to CIRCUIT proofreading (Case 78) — that case is about deploying capability against automation failure; this case is about building the capability in the first place at the edge of the trainees' prior preparation.

BACKGROUND

The recurring story in STEM workforce-development at the undergraduate level is that the pipeline narrows at every stage, and that the narrowing falls disproportionately on students whose prior preparation has not included the kind of access — to research labs, to mentorship networks, to technical-skills training that meet a research project's immediate needs — that converts undergraduate interest into research capability. CIRCUIT was built to engineer that conversion at the edge of trainees' prior preparation, across a cohort drawn from "trail-

blazing" backgrounds: first-generation college students, low-income students, and students with limited prior research access.[0]

THE INTERVENTION

The program model is the case's first contribution. Eight pillars are named and operationalized: holistic recruiting that does not screen on credentials a trailblazer's background would not have generated; mission-driven research that lets trainees see why the technical skill they are acquiring matters; targeted technical training built around the project's immediate needs; leadership development; high-resolution assessment (the assessment is the high-resolution version, not a summative pass/fail); diverse mentorship; university partnerships that route the cohort across institutions; and career empowerment that sustains the capability beyond the program. The model is published as a model in the ASEE paper — a peer-reviewed full paper with DOI, not an institutional press release.[1]

HOW IT WORKED

The longitudinal outcome evidence is the case's second contribution. The ASEE paper reports outcomes gathered over the six program cycles from 2017 through 2023 — cohort sizes growing year over year, with over 100 students supported in 2022. The program presents itself as a replicable model, with the documentation, assessment instruments, and pillar- level operationalization that a replicating institution would need.[2]

THE EVIDENCE

The honest framing preserved in the case is the one the editorial discipline demands. This is a self-authored multi-cycle program evaluation at a single program. The ASEE paper clears the peer-review bar that the build list 02 requires; the model and the operational evidence are auditable. What an external comparative evaluation — by a researcher unaffiliated with the program — would add is the causal-claim half of the evidence: did CIRCUIT produce these outcomes, or did the cohort selection produce them? The case names the open question rather than answering it.[3]

WHAT TRANSFERRED

In pair with Case 78 (CIRCUIT proofreading + MICrONS), the case completes the CIRCUIT picture: building the capability (this case, peer-reviewed) and deploying it against automation failure (Case 78, frontier with evidence-tier flag). The pair also exercises the corpus's COI discipline — both cases carry editor-related COI, both are rendered with the standing gold-bordered "Disclosure" block under the title, and both are anchored to the strongest peer-reviewed evidence available with the institutional and program-evaluation gap honestly named. The book's credibility on these cases rests on plain disclosure, not on hiding the affiliation.

> *The deliverable is the replicable model, not the program brand. The pillars are operationalized for independent evaluation.*
>
> EDITORS' SYNTHESIS OF CERVANTES ET AL. (2023).

REFERENCES

1. Cervantes, Floryanzia, Sharp, Gray-Roncal, & Johnson (2023), "Empowering Trailblazers toward Scalable, Systematized, Research-Based Workforce Development," ASEE Annual Conference, doi:10.18260/1-2–43271.
2. CIRCUIT program documentation (JHU Hub, 2017 – present) — institutional program description.
3. National Academies / NSF research on undergraduate research experience effects on STEM persistence (broader literature against which the case sits).
4. Paired case (78) — CIRCUIT proofreading + MICrONS — the deployed-capability companion.

THE LEARNING ENGINEERING LENS

LE INSIGHT

CIRCUIT is a peer-reviewed eight-pillar workforce-development program for STEM trainees from trailblazing backgrounds, with longitudinal outcomes over six program cycles. The strongest honest framing is self-authored multi-cycle program evaluation; an external comparative would add the causal half. COI under the title — editor is senior author — is binding.

LENS APPROACH

CIRCUIT workforce model is the equity-engineered workforce-development case (induced 1.2; LENS D2/PT4; with 8.3 alternate). LENS uses it in Domain 2 (Iterative Development) for the operationalized pillar model and the multi-cycle iteration evidence; in Domain 4 (Test and Evaluation) for the high-resolution assessment; and in Domain 5 (Navigating Sociotechnical Constraints) for the holistic-recruiting design choice that converts equity from rhetoric to enrolled cohort. Pair with Case 78 (CIRCUIT proofreading) — building capability vs. deploying it against automation failure. COI render is binding.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
WORKFORCE DEV STEM TRAINING EQUITY	T K	D1 **D2** D3 D4 D5 Iterative Development	4	1.2	2 · 4

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Operationalize the program pillars at the same level of detail an external replicator would need; the deliverable is not the program brand, it is the replicable model.
- Build holistic recruiting to actively de-weight credentials a trailblazer background would not have generated; the design choice converts the equity commitment from rhetoric into engineered enrollment.
- Pair high-resolution assessment with the targeted training so feedback can move at the cadence the work demands, not the semester boundary.

IN OPERATION / AFTER

- Commission external comparative evaluation alongside the self-evaluation; the causal-claim half of the evidence is what the program cannot produce on its own, and the case is honest about the gap.
- Publish the model documentation, instruments, and pillar operationalization openly so the model can be replicated and evaluated by other institutions on independent cohorts.
- Render the COI disclosure under the title — the standing language — and preserve the editorial critical distance from the program's own self-presentation in any drafting.

REFLECTION QUESTIONS

1. Identify a STEM workforce-development program in your domain. Which of CIRCUIT's eight pillars are operationalized at the level of detail an external replicator would need, and which are at the level of brand? Where would replication actually be testable?

2. Specify the external comparative evaluation you would commission alongside your own program evaluation. What independent cohort, what instrument, on what cadence — and who is unaffiliated enough to evaluate?

3. The COI render under this case's title is the standing language for editor-author cases. Identify a case in your domain where the program's evaluation is conducted by people with a stake in the program. What is the disclosure architecture that keeps the evaluation honest without hiding the affiliation?

WHO BUILDS THIS learning science & instructional design · knowledge management & data engineering — plus domain experts and a learning engineer to integrate. TOOLS scenario simulation · competency models · knowledge bases · data pipelines

CASE 72 K-12 MATHEMATICS HOMEWORK SUPPORT FORMATIVE ASSESSMENT 2014 – PRESENT

ASSISTments — National Replication and Long-Term Follow-Through

T Training Gap K Knowledge & Institutional Memory D Designed Out

IMPACT Cluster RCT across 43 schools and 2,850 students (Maine): 7th-graders assigned to ASSISTments outperformed controls on end-of-year math; largest gains for lower-performing students; a later moderator analysis (Murphy et al. 2020) reported larger minority-student gains, though fragile in the ~93%-white sample; effect persisted into 8th-grade outcomes in the 2020 follow-up

IN BRIEF

The Roschelle, Feng, Murphy, and Mason cluster RCT (AERA Open 2016), conducted across 43 schools and 2,850 students in Maine, found that 7th-graders assigned to ASSISTments outperformed controls on end-of-year mathematics, with the largest gains for lower-performing students. A later moderator analysis (Murphy et al. 2020) reported larger gains for minority students — fragile in the nearly all-white (93%) trial sample — and found the 7th-grade effect persisted into 8th-grade outcomes. A subsequent Arnold Ventures-funded extension tested a lower-cost virtual-training adaptation in predominantly rural areas, with longitudinal follow-through extended through end of 8th grade. The case is one of the few EdTech tools in the corpus with replicated multi-state RCT evidence at meaningful effect sizes and with deliberate attention to the heterogeneity that matters most for equity outcomes. Pair with Case 17 (spaced education RCTs) for the replication-discipline thread. Open questions the authors carry: whether the virtual-training adaptation matches the in-person-training arm; whether the effect persists post-grade-8.

BACKGROUND

ASSISTments is structurally different from the intelligent tutoring systems that dominate the K-12 EdTech evidence base. It augments homework rather than replacing curriculum; it does not require the institutional commitment to a new instructional system that Cognitive Tutor and its peers require; and the research team behind it (Heffernan and collaborators) has deliberately designed the platform around an evidence-generation loop with classroom teachers. The cluster RCT Roschelle et al. published in 2016 is the first national-scale evaluation of the platform, and it was designed to support claims a single-site trial could not support.[o]

THE INTERVENTION

The trial cluster-randomized 43 schools in Maine and assigned 2,850 7th-grade students to either an ASSISTments condition or a business-as-usual homework condition. The outcome instrument was end-of-year mathematics achievement. The headline finding is positive: ASSISTments-assigned students outperformed controls. The effect is meaningful in size, and the cluster randomization supports an inference at the school level rather than only at the student level. The teacher-side change required for the intervention was deliberately minimized — the platform is built around teacher-assigned homework problems, with the formative-assessment and feedback loops automated — so the trial estimates an effect achievable under realistic adoption conditions.[1]

HOW IT WORKED

The heterogeneity finding is what makes the case equity-relevant. Effect-size estimates were largest for lower-performing students. A later moderator analysis (Murphy et al. 2020) reported that minority students benefited more than the average effect would suggest — a finding that is fragile in the nearly all-white (93%) trial sample and was not an outcome the 2016 trial was powered to estimate. The pattern is the one Case 55 (Engler / enrollment algorithms) names as the inversion target: prediction and adaptive feedback used to trigger support rather than to gatekeep aid.[2]

THE EVIDENCE

Murphy et al. (2020) extended the evaluation into a longitudinal follow-through. The 7th-grade effect persisted into 8th-grade outcomes — a persistence finding that the EdTech evidence base does not consistently report, and that converts the case from a single-year effect-size study into a multi-year follow-through case. A subsequent Arnold Ventures-funded extension tested a lower-cost virtual-training adaptation in predominantly rural areas with longitudinal follow-through extended through end of 8th grade. The replication structure — trial, replication, longitudinal follow-through, adaptation tested under different deployment conditions — is the closed-loop evidence architecture the corpus's EdTech cases mostly aspire to and rarely report.[3]

WHAT TRANSFERRED

The honest open questions the case carries are the ones the research team itself names. Whether the lower-cost virtual-training adaptation matches the in-person-training arm's effect size is still under analysis. Whether the effect persists past grade 8 is the longer-horizon question that the corpus's evaluation-horizon discipline (Case 84) directly applies to. Pair the case with Case 127 (Cognitive Tutor at-scale evaluation) for the evaluation-horizon thread, with Case 17 (spaced education RCTs) for the replication-discipline thread, and with Case 55 (Engler enrollment algorithms) for the prediction-triggers-support inversion — the equity- relevant heterogeneity finding here is the structural complement to Engler's gatekeeping critique.

> *The largest gains are for lower-performing students. A later moderator analysis reports larger minority gains — but fragile in a nearly all-white sample, and not what the 2016 trial was powered to estimate.*
>
> EDITORS' SYNTHESIS OF ROSCHELLE ET AL. (2016) AND MURPHY ET AL. (2020).

REFERENCES

1. Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016), "Online Mathematics Homework Increases Student Achievement," AERA Open 2(4):1–12, doi:10.1177/2332858416673968.
2. Murphy, R. et al. (2020), follow-up evaluation extending the 7th-grade effect into 8th-grade outcomes.
3. Heffernan, N. T., & Heffernan, C. L. (2014), "The ASSISTments ecosystem," International Journal of AI in Education 24:470–497 — platform design and research-loop description.
4. Arnold Ventures RCT documentation of the virtual-training-adaptation arm — longitudinal follow-through through end of 8th grade.

THE LEARNING ENGINEERING LENS

LE INSIGHT

ASSISTments is the case in the corpus with the cleanest closed-loop evidence architecture for EdTech: cluster RCT, longitudinal follow-through into the next grade, adaptation arm under different deployment conditions, and a pre-specified equity-relevant heterogeneity finding. The case grounds the closed-loop evaluation anchor in EdTech the same way Case 40 grounds it in team-science training.

LENS APPROACH

ASSISTments is the closed-loop EdTech evaluation case (induced 2.3; LENS D2/PT4). LENS uses it in Domain 2 (Iterative Development) for the teacher-side minimum-change design discipline and in Domain 4 (Test and Evaluation) for the heterogeneity-pre-specification and longitudinal- follow-through structure. Pair with Case 84 (Cognitive Tutor at-scale evaluation horizon), Case 17 (spaced education RCTs), and Case 55 (Engler enrollment algorithms inversion — prediction-triggers-support).

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
K-12 MATHEMATICS HOMEWORK SUPPORT FORMATIVE ASSESSMENT	T K D	D1 **D2** D3 D4 D5 Iterative Development	4	2.3	2 · 4

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Pre-specify equity-relevant heterogeneity outcomes before running the trial rather than mining them afterward; the ASSISTments minority-benefit finding surfaced only in a later moderator analysis and is fragile in a nearly all-white sample — a caution about trusting subgroup effects a trial was not powered to estimate.
- Design the teacher-side change to the minimum required for the intervention to operate; the case's external-validity strength depends on its having estimated an effect achievable under realistic adoption conditions.
- Build the longitudinal follow-through into the trial's data infrastructure from the start; the 7th-to-8th-grade persistence finding required data structures that single-year trials do not necessarily provide.

IN OPERATION / AFTER

- Publish the heterogeneity result with the aggregate result; the case's equity-relevant pedagogical value depends on the heterogeneity finding being on the same page as the average effect.
- Track the adaptation arm — the lower-cost virtual-training condition — as a separate replication; the closed-loop evidence architecture the case demonstrates includes adaptation-under-different-conditions as a distinct evidence layer.
- Carry the case in syllabi alongside Case 84 so the evaluation-horizon discipline and the heterogeneity-pre-specification discipline are taught together; the two methodological lessons are independent and both load-bearing for EdTech-evaluation design.

REFLECTION QUESTIONS

1. Identify an EdTech intervention in your domain whose equity-relevant heterogeneity outcome was not pre-specified in the trial design. What pre-specification would the next replication require, and what is the data infrastructure that would support it?

2. Specify the longitudinal-follow-through design you would build into the next at-scale EdTech evaluation. What grade-to-grade or year-to-year outcome would you track, and what data infrastructure does the tracking require?

WHO BUILDS THIS learning science & instructional design · knowledge management & data engineering · systems & human-factors engineering — plus domain experts and a learning engineer to integrate. TOOLS scenario simulation · competency models · knowledge bases · data pipelines · task analysis · design prototyping

CASE 77 TUTORING HYBRID HUMAN-AI SYSTEMS K-12 EDUCATION 2024

Hybrid Human-AI Tutoring — Augmentation, Not Delegation

T Training Gap K Knowledge & Institutional Memory D Designed Out

IMPACT **Three quasi-experimental studies of hybrid human-AI tutoring deployments reported improvements in student learning relative to comparison conditions; the AI is positioned as augmentation, not delegation; the human tutor retains authorization to override and re-direct**

IN BRIEF

Thomas et al.'s LAK 2024 best paper, "Improving Student Learning with Hybrid Human-AI Tutoring: A Three-Study Quasi-Experimental Investigation," reports three quasi- experimental studies of hybrid deployments where AI augmentation is added to human tutoring rather than used to replace it. The headline finding is that learning outcomes improved relative to comparison conditions in each of the three studies. The contribution the case carries for the LENS framework is the design positioning: the AI is augmentation, the human tutor retains the authorization to override and re-direct, and the measured outcome is student learning rather than AI-system fidelity. The case is the small-tier intervention-side counterpart to Case 20 (TREWS, the clinician-AI teaming case that worked) translated into education. Pair also with Cases 78 and 68 (CIRCUIT human- AI workforce) and Case 5 (Epic Sepsis, the delegation case that did not work). Open questions: longitudinal persistence; transfer to lower-resource tutoring contexts where human-tutor availability is the binding constraint.

BACKGROUND

The deployment record for AI in tutoring has been pulled in two directions. The fully-automated tutoring track — from Cognitive Tutor through LLM-based tutoring (Case 88) — has tested whether AI alone can replace or substantially reduce the human-tutor role. The augmentation track has tested whether AI can extend the reach and effectiveness of human tutors, with the AI positioned as a tool the tutor uses rather than as a substitute for the tutor. Thomas et al.'s LAK 2024 paper is the strongest published evaluation of the augmentation track to date, and it reports three quasi-experimental studies that converge on a positive finding.[0]

THE INTERVENTION

The three studies test variants of the augmentation pattern in K-12 tutoring deployments. The AI supports the human tutor with information surfacing, problem recommendation, and student-progress visibility; the human tutor retains the conversational and pedagogical lead. Outcome measures are student learning relative to comparison conditions — students

in human-AI tutoring compared against a math-software-only baseline (Study 3 lacked a control group). Across the three studies, the hybrid condition produced measurable improvements in student learning. The replication structure across the three studies — same authorship team, varying institutional context, converging direction of effect — is the methodological backbone of the case.[1]

HOW IT WORKED

The design positioning the case carries for the LENS framework is the augmentation-not-delegation frame. The AI is positioned as augmentation, the human tutor retains the authorization to override and re-direct, and the measured outcome is student learning, not AI-system fidelity. This is the design pattern that worked in clinical-decision-support at Case 20 (TREWS) and that did not work at Case 5 (Epic Sepsis) — where TREWS preserved clinician authorization and built the explanation structure that supported it, the Epic Sepsis deployment pattern collapsed clinician judgment into alert compliance. Hybrid human-AI tutoring is the educational analog of the TREWS pattern, and the LAK 2024 paper is the evidence base that grounds the analog.[2]

THE EVIDENCE

The case anchors with the CIRCUIT pair (Cases 78 and 68) at the workforce-augmentation layer. CIRCUIT proofreading positions human capability as the recovery mechanism for automation failure at petabyte scale; the CIRCUIT workforce model builds the capability in the first place; hybrid human-AI tutoring positions AI as augmentation of an already-capable human tutor. The three cases together teach the augmentation-and-correction pattern across three deployment substrates — connectomics proofreading, neuroscience-workforce development, and tutoring — and they ground the curriculum's machine- teaming and delegation-with-revocation anchors at each substrate.[3]

WHAT TRANSFERRED

The open questions the case carries are the ones the authors name. Longitudinal effect persistence is not yet in the evidence base — the three quasi-experimental studies report end-of-intervention outcomes, not multi-year follow-through. Whether the design transfers to lower-resource tutoring contexts, where human-tutor availability is the binding constraint and the augmentation-of-a-tutor frame may not apply, is the open generalization question. The quasi-experimental design is honest about its causal-inference limits — randomization is not at the level a cluster RCT would provide — and the case carries the qualification. Future validation ongoing on persistence, transfer, and the tutor-scarce-context generalization.

> *The AI is positioned as augmentation, not delegation. The human tutor retains the authorization to override and re-direct. The measured outcome is student learning, not AI-system fidelity.*
>
> EDITORS' SYNTHESIS OF THOMAS ET AL. (2024).

REFERENCES

1. Thomas, D. R. et al. (2024), "Improving Student Learning with Hybrid Human-AI Tutoring: A Three-Study Quasi-Experimental Investigation," LAK '24, doi:10.1145/3636555.3636896.
2. Case 20 (TREWS) reference set — Henry et al. (2022), Nature Medicine — clinician-AI teaming analog.
3. Case 5 (Epic Sepsis) reference set — Wong et al. (2021), JAMA Internal Medicine — delegation-collapse analog.
4. Koedinger, K. R. et al. — Cognitive Tutor literature as the fully-automated track the augmentation track contrasts with.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Hybrid human-AI tutoring is the educational analog of the TREWS clinician-AI teaming pattern. AI is positioned as augmentation; human tutor retains override authorization; student-learning outcome is the measure. Three quasi- experimental studies converge on positive learning effects. The case pairs with Cases 20 (TREWS), 5 (Epic Sepsis), and 68 / 78 (CIRCUIT) in the human-AI teaming thread and grounds the delegation-with- revocation LEO at the educational deployment.

LENS APPROACH

Hybrid human-AI tutoring is the augmentation-not-delegation case in education (induced 6.4; LENS D3/PT6). LENS uses it in Domain 3 (Human-System Collaboration) for the augmentation pattern and the override-authorization frame, and in Domain 2 (Iterative Development) for the three-study converging-design replication. Pair with Cases 20 (TREWS) and 5 (Epic Sepsis) at the clinical analog, and with Cases 78 and 68 (CIRCUIT) at the workforce-augmentation analog.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
TUTORING HYBRID HUMAN-AI SYSTEMS K-12 EDUCATION	T K D	D1 D2 **D3** D4 D5 Human-System Collaboration	6	6.4	2 · 3

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Position the AI as augmentation explicitly in the design documentation, not implicitly in the deployment pattern; the augmentation-vs-delegation distinction is the load-bearing design choice and should be the named design choice.
- Preserve human-tutor authorization to override and re-direct as a system-design requirement, not as a discretionary affordance; the comparison with Case 5 (Epic Sepsis) is that override authorization collapses when the system pattern does not actively preserve it.
- Measure the student-learning outcome, not the AI-system-fidelity outcome; the case's pedagogical framing depends on the outcome instrument being the educationally relevant one, not the AI-development-internal one.

IN OPERATION / AFTER

- Commission longitudinal follow-through that extends the evidence base past the end-of-intervention horizon; the open persistence question is testable against the same deployment with additional data infrastructure.
- Test the augmentation design in tutor-scarce contexts; the open generalization question is whether the pattern transfers to settings where the binding constraint is human-tutor availability rather than human-tutor effectiveness.
- Pair the case with Case 20 (TREWS) in the curriculum so the augmentation-and-override pattern is taught across clinical and educational substrates; the two cases together ground the delegation-with-revocation LEO with two converging instances.

REFLECTION QUESTIONS

1. Identify an AI deployment in your domain where the design choice between augmentation and delegation has been implicit rather than explicit. What would change in the system design if the choice were named explicitly, and what comparison condition would you build to test the difference?

2. Specify the override-authorization preservation mechanism in your domain's analog deployment. Is the human operator's authority to override and re-direct a system-design requirement, a discretionary affordance, or an implicit assumption? Which of the three is honest about what the system currently supports?

WHO BUILDS THIS learning science & instructional design · knowledge management & data engineering · systems & human-factors engineering — plus domain experts and a learning engineer to integrate. TOOLS scenario simulation · competency models · knowledge bases · data pipelines · task analysis · design prototyping

CASE 80 EDUCATION AT SCALE 2012 – PRESENT

Georgia State University Predictive Analytics

T Training Gap K Knowledge & Institutional Memory

IMPACT **Six-year graduation rate 32% → 54%; equity gaps in graduation eliminated; 2,000+ more graduates per year**

IN BRIEF

Georgia State University built a predictive-analytics advising system that tracks some 800 risk factors per student and triggers proactive outreach when early warning signs appear. Crucially, it was designed with equity as a primary constraint — the explicit goal was to eliminate, not reproduce, graduation gaps — and the alerts prompt human advisors rather than making automated decisions. The six-year graduation rate rose from 32% to 54%; Black and Pell-eligible students now graduate at the same rate as their peers, and GSU produces roughly 2,000 more graduates a year. The difference from the algorithmic-bias cases (46, 49, 191) is design: GSU built equity in from the start, used predictions to trigger more human support rather than gatekeeping, and tracked outcomes by demographic group as a primary metric.

BACKGROUND

Georgia State was a regional commuter university where only about a third of students finished in six years, with large gaps by race and income. Like many institutions it had predictive data but such systems, where they existed elsewhere, were typically used for triage or gatekeeping — risking the reproduction of existing inequities rather than their repair. A model that flags at-risk students can just as easily steer them away as toward help; the same prediction serves opposite ends depending on what the institution decides to do with it.[0]

THE INTERVENTION

Beginning in 2012, GSU deployed a predictive-analytics advising system that monitors roughly 800 behavioral and academic risk factors per student daily and fires an alert to an advisor when warning signs — a missed assignment, a poor grade in a gateway course — appear. The system was built with equity as a primary design constraint, with the explicit aim of closing graduation gaps. Daily monitoring meant the alert fired while there was still time to act — a slipping student was caught at the missed assignment rather than at the failed semester, when intervention could still change the outcome.[1]

HOW IT WORKED

The load-bearing design choice was the human-loop architecture: alerts trigger proactive advising — a phone call, a meeting, a financial-aid check — rather than automated decisions. Predictions are used to deliver more support to at-risk students, not to gatekeep them out. Human judgment stays in the loop, and the model functions as decision support rather than

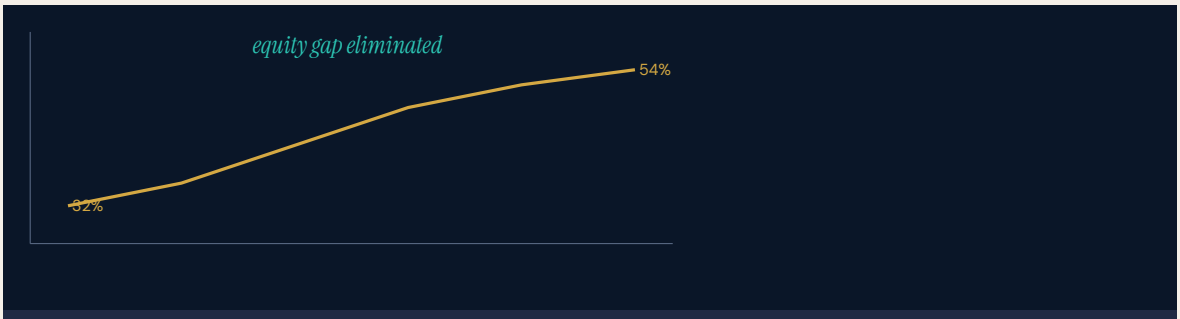
decision-maker. Routing every alert through an advisor rather than an automated action is what kept the prediction in service of the student: the model identified who needed attention, and a person decided what that attention should be.[2]

THE EVIDENCE

GSU's six-year graduation rate rose from 32% to 54%, and the institution now produces some 2,000 additional graduates a year. The graduation rate for Black students rose to match the overall rate, and Pell-eligible students graduate at the same rate as non-Pell students — the equity gap was eliminated rather than merely narrowed. Eliminating the gap rather than narrowing it is the decisive result: the overall rate rose while the disparities by race and income closed, so the gain did not come at the expense of the students the system was most at risk of leaving behind. The attribution carries a load-bearing hedge: most of the 32%-to-54% rise accumulated across a decade in which GPS Advising (launched 2012) was one component of a broader bundle of reforms — meta-majors, Panther Retention Grants, freshman learning communities, block scheduling — so the isolated causal contribution of predictive advising cannot be cleanly separated from the co-occurring changes.[3]

WHAT TRANSFERRED

GSU is the positive counterpart to the algorithmic-harm failures earlier in this part and in Part VII (the A-Level algorithm, Case 49; Robodebt, Case 191; educational predictive bias, Case 46). The same technical capability — a predictive model — produced an equity gain rather than an equity harm because of how it was framed and governed. The case is the strongest evidence that construct definition and human-loop architecture, not the model itself, determine whether prediction helps or harms. Holding the technology constant and varying only the design constraint and governance is what isolates the lesson: the same predictive capability that harmed in those cases helped here, so the framing and the human loop, not the model, are where intent lives.[4] GSU has since moved to export the model: in 2021 it founded the National Institute for Student Success, which by 2024 had advised more than 130 campuses serving some 1.5 million students, while GSU's own six-year graduation rate held steady near 53 percent — the test now shifting from whether the bundle worked at one institution to whether the advising-and-governance recipe transfers to others.[5]



GRADUATION RATE, WITH EQUITY AS A PRIMARY CONSTRAINT

> *Predictions trigger support, not gatekeeping.*
>
> EDITORS' SYNTHESIS OF THE GSU ADVISING MODEL, DRAWN FROM RENICK & STROM (2020) AND NEW YORK TIMES COVERAGE (2018)

REFERENCES

1. Renick, T. & Strom, A. (2020) on GSU's advising transformation — the system design and outcomes.
2. Georgia State University institutional research and Strategic Plan reports — graduation-rate and equity data.
3. New York Times, "How Colleges Know You're Not Finishing" (2018) — the 800-factor advising model.
4. EDUCAUSE Review on GSU predictive advising — the human-loop architecture.
5. Complete College America, Game Changers documentation — dissemination of the model.
6. Georgia State University National Institute for Student Success (founded 2021) — reported engagement with 130-plus campuses and 1.5 million students by 2024.

THE LEARNING ENGINEERING LENS

LE INSIGHT

GSU is the positive counterpart to A-Level (49), Robodebt (191), and educational algorithmic bias (46). The same technical capability — a predictive model — was deployed under a different design constraint and produced an equity outcome rather than an equity harm. The case is the strongest available evidence that the **construct definition** and the **human-loop architecture** determine whether a predictive model produces good or harm, not the model itself.

LENS APPROACH

LENS treats GSU as the canonical positive case for predictive analytics in education. LEN 4 examines the measurement architecture that made equity a primary outcome. LEN 7 examines the governance structure that kept the system as decision support rather than automated decision. LEN 1 uses it as a problem-framing exemplar.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
EDUCATION AT SCALE →	T K →	D1 D2 D3 **D4** D5	5	8.3	4
		Test and Evaluation			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Set equity as a primary design constraint from the start — the explicit aim of closing gaps — rather than discovering disparities after deployment.
- Build a human-loop architecture so alerts trigger proactive support routed through an advisor, with the model as decision support and a person deciding the action.
- Tune the monitoring to fire early — at the missed assignment, not the failed semester — so the intervention reaches the student while it can still change the outcome.

IN OPERATION / AFTER

- Track outcomes by demographic group as a primary metric, so the system is judged on whether it closes gaps rather than merely raises the average.
- Confirm the overall gain does not come at the expense of the most at-risk students — eliminating the gap, not just narrowing it, is the test that the design held.
- Keep human judgment in the loop as the system scales, so prediction continues to deliver more support rather than drifting into automated gatekeeping.

REFLECTION QUESTIONS

1. What is the difference between GSU's predictive analytics and the algorithmic-bias failure cases in Parts II and VII (e.g., Cases 46, 49, 187, 191)? Be specific about what makes the GSU implementation work.

2. Design the equity-as-primary-constraint version of a predictive system in your domain. What would you measure first?

3. GSU used predictions to deliver more support rather than to gatekeep, with an advisor between the alert and the action. Identify a predictive system in your domain and specify the human-loop architecture that would keep it serving the people it flags rather than screening them out.

WHO BUILDS THIS learning science & instructional design · knowledge management & data engineering — plus domain experts and a learning engineer to integrate. TOOLS scenario simulation · competency models · knowledge bases · data pipelines

CASE 93 WORKFORCE DEV NATIONAL L&D POLICY ASIA-PACIFIC 2015 – PRESENT

Singapore SkillsFuture — National Workforce Capability at Scale

G Governance & Trust K Knowledge & Institutional Memory D Designed Out

EVIDENCE TIER practice-synthesis-tier — source confidence flagged; future validation ongoing.

IMPACT **Singapore's SkillsFuture pairs individual training credits with employer subsidies, a cross-sector skills framework, and a two-wave outcome survey (TRAQOM, at end-of-course and at six months) — a 2018 MTI study found a 5.8% real wage premium for WSQ-trained workers, with 87% of Work-Study Programme graduates employed full-time within six months**

IN BRIEF

Singapore's SkillsFuture Movement, launched in 2015, pairs individual training credits with employer subsidies, a cross-sector skills framework, and one of the most ambitious national L&D measurement instruments deployed at scale: the two-wave TRAQOM survey, administered at end-of-course and at six months post-training, paired with labor-market data. Self-reported figures from the 2024 Year-in-Review are strong: 98% of trainees report being able to perform better at work; 93% report the course played a pivotal role. A separate Work-Study Programme Outcomes Survey found 87% of graduates employed full-time within six months (2019 cohort; the 2024 review reports more than 9 in 10). A 2018 MTI study found a 5.8% real wage premium for WSQ-trained workers. The honest reading the case carries into print: self-report dominates the headline numbers, and the program has not been subjected to a rigorous quasi-experimental external evaluation that would isolate the program's causal effect from underlying labor-market trends. SkillsFuture is the non-US national-scale L&D case the corpus needs for both the corporate / workforce L&D gap and the non-US/UK/EU gap. The evidence-tier flag is practice-synthesis: the program design and the TRAQOM instrument are documented in SSG annual reports and in ILO and Springer analyses, the headline outcomes are self-report, and future validation — particularly quasi-experimental causal evaluation — is ongoing.

BACKGROUND

SkillsFuture was launched in 2015 as a Singapore-wide workforce-capability program at the seam between individual upskilling, employer demand, and state coordination. The program design pairs individual training credits (SkillsFuture Credit), employer subsidies for work-force training, a cross-sector skills framework that defines competencies and progression paths across industries, and a Work-Study Programme that embeds learners in employer contexts during training.[o]

THE INTERVENTION

The measurement instrument is unusually ambitious for a national L&D program. The Training Quality and Outcomes Measurement framework (TRAQOM) is a two-wave outcome survey administered at end-of-course and at six months post-training. It is paired with labor-market data so that self-reported outcomes can be cross-checked against employment and wage outcomes at population scale. The design crosses the Kirkpatrick Level-2 / Level-3 seam (Case 79) at policy level, not only program level.[1]

HOW IT WORKED

The 2024 Year-in-Review reports headline figures: 98% of trainees report being able to perform better at work; 93% report the course played a pivotal role. A separate Work-Study Programme Outcomes Survey found 87% of graduates employed full-time within six months (2019 cohort; the 2024 review reports more than 9 in 10 employed). A 2018 study by the Ministry of Trade and Industry found a 5.8% real wage premium for workers with a Workforce Skills Qualifications (WSQ) certification. The labor-market figures are the strongest available external corroboration of the self-report data.[2] Since 2024 the program has broadened well beyond credits and surveys: the SkillsFuture Level-Up Programme gave Singaporeans aged 40 and over a S$4,000 credit top-up (May 2024) and, from 2025, a training allowance of up to S$3,000 a month for full-time reskilling, while the SkillsFuture Jobseeker Support scheme (launched April 2025) added up to S$6,000 over six months for the involuntarily unemployed — extending the model from upskilling into income-supported labour-market transition.

THE EVIDENCE

The honest reading is the load-bearing teaching point. Self-report dominates the headline outcomes. The program has not been subjected to a rigorous quasi-experimental external evaluation that would isolate the program's causal effect from underlying labor-market trends — and Singapore's labor market has been strong across the program's deployment period. The TRAQOM design is one of the strongest national L&D instruments deployed, and what it cannot yet do is what no national L&D instrument yet does well: produce decision-grade causal evidence at population scale. Future validation is ongoing.[3]

WHAT TRANSFERRED

The LENS teaching point is that the program is a non-US national-scale case for the corporate / workforce L&D cluster (Cases 79, 65, 83, 70) and a non-US/UK/EU case for the geographic-coverage gap. The amended LEO on collaboration measurement is directly exercised: TRAQOM measures across employer-employee-state, not only across the training organization. Pair with Case 18 (PEPFAR) for the global-health workforce-capability counterpart, and with Case 70 (HILS) for the design-side practitioner pattern that the SSG program operationalizes at policy scale. Evidence-tier flag is practice-synthesis; the design is documented, the causal magnitudes are open.[4]

> *The instrument crosses the Level-2 / Level-3 seam at policy level. What it cannot yet do is what no national L&D instrument yet does well.*
>
> EDITORS' SYNTHESIS OF THE SKILLSFUTURE MOVEMENT AND THE TRAQOM MEASUREMENT FRAMEWORK.

REFERENCES

1. SkillsFuture Singapore (SSG), Year-in-Review 2024 — program metrics and outcome reporting.
2. Ministry of Education (MOE), Singapore, parliamentary replies on TRAQOM, 2020.
3. International Labour Organization (ILO), "Investigating an Up-skilling Programme in Singapore" — international comparative analysis.
4. "Future-Skilling the Workforce: SkillsFuture Movement in Singapore," Springer, 2024 — peer-reviewed program analysis.
5. Ministry of Trade and Industry (MTI), Singapore, 2018 — WSQ wage-premium study.

THE LEARNING ENGINEERING LENS

LE INSIGHT

SkillsFuture is the non-US national-scale L&D case the corpus needs. The TRAQOM instrument is among the most ambitious national L&D measurement architectures deployed; the headline outcomes are self-report dominant; no rigorous quasi-experimental external evaluation yet exists. Evidence-tier flag is practice-synthesis; future validation is ongoing.

LENS APPROACH

SkillsFuture is the national workforce-capability case (induced 2.4; LENS D5/PT4). LENS uses it in Domain 5 (Navigating Sociotechnical Constraints) for the amended LEO on collaboration measurement — TRAQOM measures across employer-employee-state — and in Domain 2 as the policy-scale operationalization of the HILS-style environment-and-event integration (Case 70). Pairs with Case 18 (PEPFAR) for the global-health workforce-capability counterpart.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
WORKFORCE DEV NATIONAL L&D POLICY ASIA-PACIFIC	G K D	D1 D2 D3 D4 → D5 Navigating Sociotechnical Constraints	4	2.4	2 · 5 · 3

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Design the measurement instrument across the training-and-employment seam, not within the training organization alone — TRAQOM's two-wave + labor-market cross-check is the architecture the LENS Domain 4 module should teach.
- Make the cross-sector skills framework a first-class artifact — without it the credits, the subsidies, and the Work-Study Programme do not cohere as a single workforce-capability deliverable.
- Treat the self-report dominance honestly: name what TRAQOM can and cannot establish at the design stage, so the program documentation does not have to retrofit the hedge.

IN OPERATION / AFTER

- Pair with Case 18 (PEPFAR) for the global-health workforce-capability counterpart at multi-country scale; together they teach what national- and program-scale L&D measurement at evidence-flagged tier looks like.
- Use the amended LEO on collaboration measurement: TRAQOM is a worked example of measurement across employer-employee-state, and the program documentation can teach the architecture in LENS Domain 5 (Sociotechnical Constraints).
- Carry the practice-synthesis flag honestly: the program design and the TRAQOM instrument are documented, the headline magnitudes are self-report, and future validation requires independent quasi-experimental causal evaluation.

REFLECTION QUESTIONS

1. Identify a workforce-capability program in your context that currently measures at the training-organization boundary. What would the analogue of TRAQOM — a two-wave outcome survey paired with employment-and-wage data at population scale — require of your measurement infrastructure?

2. Specify the cross-sector skills framework that would coordinate individual credits, employer subsidies, and a Work-Study-style placement program in your context. Without the framework, do the components cohere as a single capability deliverable?

3. SkillsFuture's headline magnitudes are self-report dominant. What independent quasi-experimental evidence — comparison-cohort design, regression discontinuity, instrumented variation — would you require before treating any specific outcome magnitude as decision-grade for a program-scale investment in your context?

WHO BUILDS THIS governance, ethics & measurement · knowledge management & data engineering · systems & human-factors engineering — plus domain experts and a learning engineer to integrate. TOOLS governance frameworks · bias & evidence audits · knowledge bases · data pipelines · task analysis · design prototyping

CASE 95 K-12 EDUCATION SCHOOL CLIMATE IMPLEMENTATION FIDELITY 1997–2024

PBIS Implementation Fidelity — Why the Framework Travels Only with Its Measurement Loop

T Training Gap G Governance & Trust

DISCLOSURE Institutional overlap: an editor shares an institution (Johns Hopkins School of Education) with leading PBIS researchers; no editor was personally involved. Framed as learning from that peer-reviewed literature.

IMPACT A multi-tiered behavior-support framework implemented in more than 25,000 US schools; group-randomized effectiveness trials in 37 Maryland elementary schools found significant reductions in suspensions and office discipline referrals and improvements in child behavior problems when the model was implemented with fidelity; the PBIS Maryland state-university-district partnership (since 1999) trained staff at over 1,000 schools around a standing fidelity-measurement and coaching loop

IN BRIEF

Positive Behavioral Interventions and Supports (PBIS) is one of the most widely scaled evidence-based frameworks in US education — implemented, by the Center on PBIS's own count, in more than 25,000 schools. The evidence base is unusually strong for education at scale: Bradshaw, Mitchell, and Leaf's five-year group-randomized effectiveness trial in 37 Maryland elementary schools (2010) found that schools trained in school-wide PBIS reached high implementation fidelity and experienced significant reductions in suspensions and office discipline referrals, with companion trials reporting improvements in child behavior problems (Bradshaw, Waasdorp, & Leaf, Pediatrics 2012) and reductions in bullying and peer rejection (Waasdorp et al. 2012). The central finding the implementation-science literature itself reports — and the reason the case is in this book — is that the effect does not live in the framework document. It lives at the interface between the intervention design and the school's implementation capability: fidelity measurement (the School-wide Evaluation Tool, the Tiered Fidelity Inventory), coaching infrastructure, district capacity, and sustained administrative support carry the effect, and schools that adopt the framework without that layer show attenuated or null results. The Maryland scale-up — PBIS Maryland, a state-university-district partnership running since 1999 — is the worked example of scaling **with** the fidelity loop attached. The hedges are the literature's own: fidelity varies widely, sustainability is fragile, and abandonment of effectively implemented practice is commonplace.

BACKGROUND

The recurring story in school-improvement research is that an intervention validated in trials arrives at a new school as a document — a framework, a binder, a training day — and produces nothing. PBIS is the instructive counter-example because its research community confronted that problem directly and instrumented it. The framework itself, built from applied behavior analysis and organized as a multi-tiered system of support (universal expectations and reinforcement at Tier 1, targeted group supports at Tier 2, individualized supports at Tier 3), was codified in the late 1990s under an OSEP technical-assistance center and has since scaled to more than 25,000 US schools. The question the field then spent two decades answering is the one this casebook cares about: why does the same framework work in some schools and not others?[0]

THE INTERVENTION

The causal evidence comes principally from the Maryland group-randomized effectiveness trials. Bradshaw, Mitchell, and Leaf (2010) followed 37 Maryland elementary schools over five years; schools randomized to school-wide PBIS training implemented the model with high fidelity and showed significant reductions in student suspensions and office discipline referrals relative to comparison schools. Companion analyses from the same trial infrastructure reported significant improvements in teacher-rated child behavior problems across 12,344 children (Bradshaw, Waasdorp, & Leaf 2012) and reductions in bullying and peer rejection (Waasdorp, Bradshaw, & Leaf 2012). The load-bearing qualifier is in the trial design itself: the randomized schools did not merely receive the framework — they received structured training, on-site coaching, and annual fidelity assessment, and the published effects are effects of that package implemented with fidelity.[1]

HOW IT WORKED

The fidelity instruments are the part of the story this book reads as capability measurement. The School-wide Evaluation Tool (SET; Horner et al. 2004) and its successor, the Tiered Fidelity Inventory (TFI; Algozzine et al. 2014, with technical-adequacy validation in McIntosh et al. 2017), turn "is the school actually doing PBIS?" into a scored, repeatable observation — expectations defined and taught, reinforcement systems operating, data-based decision cycles running, administrative leadership engaged. The instruments are not compliance paperwork; they are the closed loop that links the intervention to its evidence. Fidelity scores feed the coaching conversation, coaching moves the scores, and the outcome literature consistently reports that schools implementing below fidelity thresholds — typically those adopting the framework without the coaching and measurement layer — show attenuated or null effects relative to high-fidelity implementers.[2]

THE EVIDENCE

The Maryland scale-up is the worked example of carrying that loop to state scale. PBIS Maryland, formed in 1999 as a partnership among the Maryland State Department of Education, Sheppard Pratt Health System, and Johns Hopkins University, with all 24 local school systems as partners, has trained staff at over 1,000 Maryland schools. The design decision that distinguishes it from framework-by-memo scale-ups is structural: each district provides local coaching capacity, the state provides training and annual fidelity measurement, and the university partners run the prevention research on the resulting data — including the randomized trials above and a subsequent statewide quasi-experimental study of the

scale-up itself. The scale-up did not distribute a document; it replicated the implementation infrastructure that the trials had shown to carry the effect.[3]

WHAT TRANSFERRED

The honest framing is the one the literature itself supplies, and the case preserves it as learning rather than critique. Sustainability is the field's own named open problem: McIntosh and colleagues, analyzing fidelity data from roughly 3,000 schools, found that speed of initial fidelity attainment predicts sustained implementation, that middle and high schools are at elevated risk of low implementation, and that abandonment of effectively implemented practice is commonplace (McIntosh et al. 2013, 2016). Sugai and Horner's 2020 synthesis frames sustaining and scaling as a distinct engineering problem from demonstrating efficacy. That the PBIS research community measured its own implementation variance, published its own null-attenuation conditions, and built the instruments that make the variance visible is precisely why the case sits in the intervention corpus: the field solved — in the open, with peer review — the coupling problem that most deployments in this book never instrumented at all.[4]

> *The framework document does not carry the effect. The fidelity loop — measurement, coaching, district capacity, sustained administrative support — is the intervention.*
>
> EDITORS' SYNTHESIS OF THE SWPBIS IMPLEMENTATION LITERATURE.

REFERENCES

1. Bradshaw, Mitchell, & Leaf (2010), "Examining the Effects of Schoolwide Positive Behavioral Interventions and Supports on Student Outcomes: Results From a Randomized Controlled Effectiveness Trial in Elementary Schools," Journal of Positive Behavior Interventions, 12(3), 133 – 148.
2. Bradshaw, Waasdorp, & Leaf (2012), "Effects of School-Wide Positive Behavioral Interventions and Supports on Child Behavior Problems," Pediatrics, 130(5), e1136 – e1145.
3. Horner, Todd, Lewis-Palmer, Irvin, Sugai, & Boland (2004), "The School-Wide Evaluation Tool (SET): A Research Instrument for Assessing School-Wide Positive Behavior Support," Journal of Positive Behavior Interventions, 6(1), 3 – 12.
4. McIntosh, Massar, Algozzine, George, Horner, Lewis, & Swain-Bradway (2017), "Technical Adequacy of the SWPBIS Tiered Fidelity Inventory," Journal of Positive Behavior Interventions, 19(1), 3 – 13.
5. McIntosh, Mercer, Nese, Strickland-Cohen, & Hoselton (2016), "Predictors of Sustained Implementation of School-Wide Positive Behavioral Interventions and Supports," Journal of Positive Behavior Interventions, 18(4), 209 – 218.
6. Bradshaw, Pas, et al. (2012), "A State-Wide Partnership to Promote Safe and Supportive Schools: The PBIS Maryland Initiative," Administration and Policy in Mental Health, 39(4), 225 – 237.

THE LEARNING ENGINEERING LENS

LE INSIGHT

PBIS is the rare education framework whose research community instrumented its own implementation variance: randomized trials showed real effects on suspensions, referrals, and behavior — conditional on fidelity — and the field built the measurement instruments (SET, TFI), the coaching infrastructure, and the sustainability literature that explain why the same framework works in one school and not another. The effect lives at the coupling between design and implementation capability, and the fidelity instruments are the capability measurement layer.

LENS APPROACH

PBIS implementation fidelity is the coupling-and-fidelity intervention case (induced 2.3; LENS D2/PT3). LENS uses it in Domain 2 (Iterative Development) for the fidelity-measurement-to-coaching loop as the mechanism that carries an evidence-based design into practice; in Domain 4 (Test and Evaluation) for the SET/TFI instruments as validated measures of implementation rather than outcome; and in Domain 5 (Navigating Sociotechnical Constraints) for the Maryland state-university-district partnership as the governance architecture that made a 1,000-school scale-up sustain its measurement discipline. The framing is learning

from the implementation-science literature — the field documented its own attenuation conditions — and the COI render under the title is binding.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
K-12 EDUCATION SCHOOL CLIMATE IMPLEMENTATION FIDELITY	T G	D1 **D2** D3 D4 D5 Iterative Development	3	2.3	2

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Treat the fidelity instrument as part of the intervention, not as evaluation overhead; a framework shipped without its measurement loop is a different — and weaker — intervention than the one the trials validated.
- Build the coaching layer at the district, not just the school; the Maryland design routes coaching capacity through every partner district so a principal transition does not orphan the implementation.
- Score fidelity on a public, repeatable rubric (SET/TFI-style) and feed the scores into the coaching cadence, so the implementation conversation runs on observations rather than impressions.

IN OPERATION / AFTER

- When an evidence-based practice underperforms at a new site, attribute the gap with fidelity data before revising the practice; the PBIS literature shows most of the variance lives in implementation, not design.
- Plan for sustainability as its own engineering problem: instrument speed-to-fidelity, watch the elevated-risk tiers (middle and high schools, in this literature), and treat abandonment as a measurable failure mode rather than an anomaly.
- Publish the implementation variance, including the null and attenuated conditions; the PBIS field's credibility rests on having documented where its own framework does not work.

REFLECTION QUESTIONS

1. Pick an evidence-based practice in your domain that underperformed at a new site. What would its SET or TFI look like — the scored, repeatable observation that distinguishes "the practice failed" from "the practice was never implemented"? Who would administer it, and on what cadence?

2. The Maryland scale-up replicated infrastructure — district coaching capacity, annual fidelity measurement, a university research partner — rather than distributing a framework document. Map your own scale-up plan against those three layers. Which one is missing, and what does the PBIS sustainability literature predict happens without it?

3. McIntosh and colleagues found abandonment of effectively implemented practice is commonplace, and that speed of initial fidelity attainment predicts sustainability. What early-implementation signal plays that role in your setting, and is anyone measuring it?

WHO BUILDS THIS learning science & instructional design · governance, ethics & measurement — plus domain experts and a learning engineer to integrate. TOOLS scenario simulation · competency models · governance frameworks · bias & evidence audits

Chapter 5

Aviation & Aerospace — What Fails

When automation, transition, and assumption meet an unforgiving envelope.

The aircraft did what it was designed to do. So did the crew. The designs disagreed.

AN OBSERVATION RECURRING ACROSS THE CHAPTER'S CASES.

CASE 97 AVIATION 2018 – 2019

Boeing 737 MAX / MCAS

D Designed Out T Training Gap H Human-System Interface

IMPACT 346 killed across two crashes — Lion Air 610 and Ethiopian Airlines 302; 20-month worldwide grounding; ~\$20B direct cost; FAA delegated-authority reform under the Aircraft Certification, Safety, and Accountability Act (2020)

IN BRIEF

The Boeing 737 MAX was a re-engined 737 meant to fly without new pilot training — the commercial promise that sold it against the Airbus A320neo. Its bigger, forward-mounted engines changed the jet's pitch behavior, so Boeing added software (MCAS) to mask the difference, then kept it out of the manuals and training and let it fire repeatedly on a single angle-of-attack sensor. When that sensor failed on Lion Air 610 (October 2018) and Ethiopian 302 (March 2019), MCAS forced the nose down and crews who had never heard of it could not recover; 346 people died and the fleet was grounded for twenty months. Five major investigations — the NTSB-supported KNKT (Lion Air) and EAIB (Ethiopian) reports, the US House Transportation Committee final report, the DOT Inspector General review, and the multinational Joint Authorities Technical Review (JATR) — converged on the same mechanism: an MCAS that could fire the nose down on a single failing sensor, waved through an FAA delegated-authority regime in which Boeing reviewed much of its own safety judgment. The House Transportation Committee's report went further, documenting the training omission as a cost-driven commercial choice — Boeing had promised Southwest a $1-million-per-plane rebate if simulator training proved necessary. The MAX is the book's inverse case: human capability engineered out of a system to save the price of sustaining it, with the elision underwritten by a certification process that did not have the independence to catch it.

BACKGROUND

The 737 MAX was Boeing's answer to the Airbus A320neo: a re-engined version of its best-seller that airlines could fly without retraining pilots on a new type. But the MAX's larger, forward-mounted engines changed how it pitched at high angles of attack, so Boeing added software — the Maneuvering Characteristics Augmentation System, MCAS — to make it handle like the older 737 NG, papering over an aerodynamic change with a control law so the airframe would feel identical from the cockpit. The whole commercial logic depended on that software staying invisible: the less of a "new function" MCAS seemed, the less the MAX would trigger costly new training, and the lower the airline's switching cost away from the A320neo. Boeing reportedly promised Southwest a rebate of about a million dollars per jet if simulator training proved necessary — a clause that turned the training requirement into a direct line on the program's ledger, and one the House investigation later cited as a structural reason the "no new training" promise behaved as a binding constraint on the engineering rather than as an aspiration.[0]

WHAT HAPPENED

To keep MCAS in the background, it was left out of the manual and the type-rating training, and allowed to fire on a single angle-of-attack sensor with no second source to cross-check it. The system's authority was also expanded late in development — the trim it could command grew from 0.6° to 2.5° per cycle, and its firing was made repeatable rather than one-shot — without the corresponding reassessment of the failure modes that change implied. When the sensor failed on Lion Air Flight 610 in October 2018, MCAS repeatedly forced the nose down; the crew, never told the system existed, fought it cycle after cycle until the jet dove into the Java Sea, killing all 189. Five months later Ethiopian Airlines Flight 302 repeated almost exactly the same sequence — sensor failure, repeated trim commands, unrecoverable nose-down attitude — killing 157, for 346 dead across the two crashes, and the entire MAX fleet grounded worldwide for what would become twenty months.[1]

THE INVESTIGATION

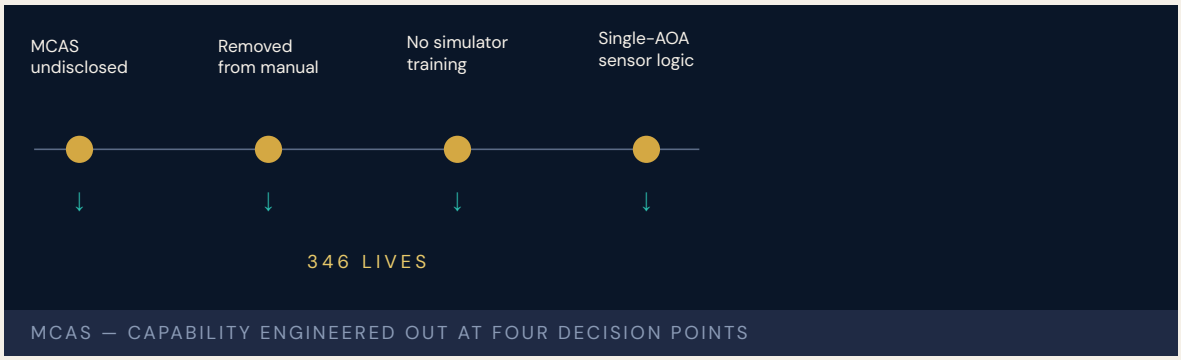
The investigations made the decisions legible. The Indonesian KNKT and Ethiopian EAIB accident reports, with NTSB participation, established the accident sequences and the single-sensor design as the proximate cause. The multinational Joint Authorities Technical Review (JATR), convened by the FAA in 2019 with international regulators, concluded that MCAS had not been evaluated as the novel flight-control function it actually was. Internal 2013 Boeing meeting minutes showed employees noting that calling MCAS a new function would bring "greater certification and training impact" — the cost the program was built to avoid, written down years before either crash; a 2016 internal survey found 39% of certification staff felt undue management pressure. The House Transportation Committee's 2020 final report concluded Boeing's assumption that simulator training was unnecessary "diminished safety, minimized the value of pilot training, and inhibited technical design improvements," naming the omission as a choice rather than an oversight.[2] The DOT Inspector General traced how the single-sensor design and the training omission passed through a certification process in which the FAA had delegated much of the safety judgment back to Boeing itself under the Organization Designation Authorization (ODA) program — so the company effectively reviewed its own most consequential trade-offs, and the regulator lacked both the technical depth and the institutional independence to challenge the assessment.[3]

THE CAPABILITY GAP

The MAX inverts this book's usual case. Human capability was not overlooked by accident; it was deliberately engineered out to avoid the cost of the training that would have created it. The pilots were not undertrained by oversight but by design — the absence of training was, in effect, a contract deliverable promised to customers and protected by a rebate. The single-sensor architecture was a second engineered-out capability: a redundant cross-check on the angle-of-attack signal would have triggered the kind of design and certification work the program was built to avoid, and was left out for the same reason. Seen together, the crashes are not a training failure that befell a good airplane but exactly what the commercial and engineering decisions specified: a flight-control system that depended on pilots reacting correctly, in seconds, to a failure they had been guaranteed never to learn about, with no sensor cross-check that could keep the failure from arriving in the first place.[4]

AFTERMATH & REFORM

The MAX was grounded for some twenty months — the longest grounding of a US-certified airliner in the jet era. MCAS was redesigned to use both AOA sensors, to fire once rather than repeatedly, and with limited authority bounded by airspeed — restoring the margins the original had stripped away. The FAA's 2020 return-to-service directive required the simulator training the airplane had been built to avoid. The Aircraft Certification, Safety, and Accountability Act of 2020 then tightened FAA oversight of the ODA delegation regime, required disclosure to pilots of flight-control systems that act on a single sensor input, and funded FAA engineering capacity that the delegation system had let atrophy. Boeing entered a Deferred Prosecution Agreement with the DOJ on a fraud charge tied to its disclosures to the FAA's Aircraft Evaluation Group.[5] That agreement's oversight was tested when a door plug blew off a separate 737 MAX 9 (Alaska Airlines 1282) in January 2024, days before the DPA's probation period lapsed; the DOJ found Boeing had breached the agreement, a 2024 guilty plea was rejected by the court, and in May 2025 the parties settled on a non-prosecution agreement worth more than $1.1 billion — no criminal conviction, over the objection of the crash victims' families. The reform conceded the point the program had spent years resisting: the training was a real requirement all along, the sensor cross-check was a real requirement all along, and removing them from the paperwork only deferred the cost — and then moved it onto two airplanes full of people.



MCAS — CAPABILITY ENGINEERED OUT AT FOUR DECISION POINTS

> *Boeing's assumption that simulator training was unnecessary diminished safety, minimized the value of pilot training, and inhibited technical design improvements.*
>
> HOUSE TRANSPORTATION COMMITTEE REPORT, 2020

REFERENCES

1. U.S. House Committee on Transportation and Infrastructure, The Boeing 737 MAX Aircraft: Costs, Consequences, and Lessons from Its Design, Development, and Certification (Final Committee Report, Sept. 2020) — the MCAS omission, the 2013 minutes, the 2016 survey, and the "diminished safety" conclusion.
2. U.S. House report (2020) and MCAS design summary — MCAS firing on a single angle-of-attack sensor; the Lion Air 610 (189 killed) and Ethiopian 302 (157 killed) accident sequences.
3. U.S. House Transportation Committee report (2020) — Boeing internal communications and the certification-and-training-impact reasoning (quoted).
4. U.S. Department of Transportation, Office of Inspector General, correspondence and review of FAA oversight of MCAS and the angle-of-attack systems (2019–2020) — the delegated-certification process.
5. J. Herkert, J. Borenstein & K. Miller, "The Boeing 737 MAX: Lessons for Engineering Ethics," Science and Engineering Ethics 26 (2020) — certification-by-omission as an engineering-ethics failure.
6. Federal Aviation Administration, Summary of the FAA's Review of the Boeing 737 MAX (Return-to-Service, 2020) — the MCAS redesign and the simulator-training requirement; and the Aircraft Certification, Safety, and Accountability Act (2020).

THE LEARNING ENGINEERING LENS

LE INSIGHT

The 737 MAX is the documentary record of a design choice to remove human capability to avoid the regulatory and commercial cost of sustaining it. The pilots were not undertrained by accident — they were undertrained as a **requirement**. The omission of training was a contract deliverable. Once that is understood, the accident becomes legible as an engineering decision rather than a training failure.

LENS APPROACH

The 737 MAX is the canonical engineered-out-capability failure (induced 1.1; LENS D1+D3/PT4). LENS uses it in Domain 1 (Systems Analysis) for the LEO-1 work of decomposing system performance requirements into measurable human-capability requirements: the elided pilot-training requirement is the traceable artifact a capability-requirements analysis would have surfaced before the omission could become a contract deliverable. LENS uses it in Domain 5 (Navigating Sociotechnical Constraints; LEO-5) for the delegated-authority regulatory regime in which Boeing reviewed its own most consequential safety judgments — the case is the governance counterpart to Therac-25 (Case 1) at the safeguard-removed-with-nothing-in-its-place layer, and pairs with Patriot/Dhahran (Case 129) at the layer of design assumptions that do not travel with a change of context.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
AVIATION →	D T H →	D1 D2 D3 D4 D5	4	1.1	1 · 5
		Systems Analysis + D5			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Treat any control law that masks a handling change as a new function in its own right — document it, surface it to pilots, and size its training as part of the design, not after it.
- Forbid safety-critical authority from resting on a single sensor; require an independent source or a cross-check before software can command the flight surfaces.
- Quarantine commercial commitments — rebates, "no new training" promises — from the engineering judgment about what capability the airplane actually requires.

IN OPERATION / AFTER

- Audit delegated certification so the party making a trade-off is never the same party that signs off on its safety, restoring independent review of the riskiest decisions.
- Monitor in-service AOA-disagree and uncommanded-trim events as leading indicators, with a route that reaches engineering before a second airframe is lost.
- Re-validate the "no new training needed" claim against real fleet incidents, and reopen the training requirement the moment operational data contradicts the assumption.

REFLECTION QUESTIONS

1. If a customer contract required removing a training requirement you judged necessary, what artifact would you produce, who would you route it to, and what would you do if the routing failed?
2. Reconstruct the MAX accident as a **capability-requirements** failure: what was the elided requirement, at what life-cycle stage was it elided, and who had the authority to insert it?
3. MCAS was permitted to act on a single sensor with no cross-check. Identify a system you work on that takes a safety-critical action from one unverified input, and specify the redundancy or human confirmation it lacks.

WHO BUILDS THIS systems & human-factors engineering · learning science & instructional design · human factors & interaction design — plus domain experts and a learning engineer to integrate. TOOLS task analysis · design prototyping · scenario simulation · competency models · usability testing · cognitive walkthroughs

CASE 102 AVIATION 2009

Air France Flight 447

T Training Gap H Human-System Interface

IMPACT 228 killed; the deadliest accident in Air France's history

IN BRIEF

Air France Flight 447, an Airbus A330, fell into the South Atlantic on 1 June 2009, killing all 228 aboard — the deadliest accident in the airline's history. At cruise altitude the pitot probes iced over, the airspeed readings failed, and the autopilot handed the jet to a crew that had never trained to fly it by hand in that regime. The pilot flying held the nose up into a full aerodynamic stall and never recognized it; the aircraft descended for nearly four and a half minutes. The BEA traced the loss to the airspeed failure, the crew's inappropriate inputs, and a training system that taught stall prevention at low altitude but never stall recovery at altitude — a gap between the trained envelope and the operational one that reshaped global pilot-training rules.

BACKGROUND

AF447 left Rio for Paris on 31 May 2009 with 228 aboard an Airbus A330 — a highly automated jet with an excellent safety record. Its route crossed the equatorial storm band, and it carried a known vulnerability: Thales AA pitot probes prone to brief icing at high altitude in heavy precipitation, a pattern documented across the A330/A340 fleet in the years before the accident. A replacement program with newer probes was underway, but the accident aircraft had not yet been modified, so the vulnerability the program existed to close was still live on the very jet that crossed the storm band — the retrofit recognized as necessary but not yet fitted where it mattered.[0] The A330's fly-by-wire architecture also carried a design choice that would matter later: the two sidesticks were not mechanically linked, so an input on one was not felt on the other, a philosophy that traded the visual cue of yoke movement for sidestick lightness and independence.

WHAT HAPPENED

At 35,000 feet the probes iced, the airspeed readings went invalid, and the autopilot and autothrust disconnected into Alternate Law — a degraded control regime in which stall protection no longer held. The pilot flying responded with sustained nose-up input; the jet climbed, stalled, and never recovered, falling some 38,000 feet into the ocean in about four and a half minutes. The stall warning sounded, then cut out at extreme angle of attack and resumed when the nose dropped — warning against the one input that would have begun a recovery, so that the cue meant to guide the crew instead punished the correct action and rewarded the fatal one, an inversion no amount of hand-flying instinct could resolve in the time available.[1] The other pilot's correcting inputs on his own sidestick went unfelt on

the flying pilot's: the design's clean independence became, under stress, an inability to see what the other crew member was doing.

THE INVESTIGATION

The BEA's three-year investigation, concluded in the 2012 final report, named a chain: airspeed loss from pitot icing, inappropriate crew inputs, and the crew's failure to recognize and recover from the stall.[2] The pilots were not incompetent — they were outside anything their training had prepared them for, and the BEA said so: "the conditions under which pilots are trained and exposed to stalls… did not result in reasonably reliable expected behaviour patterns."[3] The finding reframed the loss from a question of individual airmanship to one of training design: the crew had been drilled in a regime that never produced the responses the emergency demanded. The BEA report also surfaced the longer industry debate over fly-by-wire philosophies — the Airbus convention of independent sidesticks and protected envelopes versus the Boeing convention of linked yokes and trim feedback — not to adjudicate which was safer in steady state but to argue that whichever a manufacturer chose, the training had to make the degraded-mode behavior of that choice reflexive rather than novel.

THE CAPABILITY GAP

The gap was precise. Airlines trained stall recovery at low altitude — the regime certification required and the one simulators could reproduce — while the simulators of the era could not faithfully reproduce a high-altitude stall, so crews never practiced the situation that arrived. Years of reliable automation had also let hand-flying skills atrophy, so when the autopilot handed back control the crew met an unfamiliar regime with rusty manual skills at the worst possible moment. The trained envelope and the operational envelope had quietly diverged, and AF447 fell into the gap between them — a gap no one had been positioned to see widen, because the training record showed full compliance and the operational record showed an excellent safety performance, right up until the regime that neither had covered arrived together with degraded control law.[4]

AFTERMATH & REFORM

The BEA's recommendations reached far beyond one airline: the pitot probes were replaced fleet-wide, and the report pressed for training in manual high-altitude flight, stall recovery from cruise altitude, unreliable-airspeed handling, and angle-of-attack indication.[5] Regulators then made Upset Prevention and Recovery Training mandatory for airline pilots — FAA Part 121 adopted UPRT-aligned stall-recovery training in 2014; EASA phased UPRT into ATPL training by 2019; ICAO codified it in Doc 10011 — closing at the regulatory level the gap that had been invisible at the airline level, and moving the fix from a single carrier's discretion to a binding standard every airline had to meet.[6] Simulator manufacturers were pushed to extend their aerodynamic models into the post-stall regime so the training could actually take place. The crew performed exactly as trained; the training was the wrong training, and only a system-wide mandate could keep that mismatch from recurring elsewhere.

54 SECONDS: STALL WARNING SOUNDED, RESPONSE NEVER CAME

> *The crew never understood that they were stalling and consequently never applied a recovery manoeuvre.*
>
> BEA, FINAL REPORT ON AIR FRANCE FLIGHT 447, JULY 2012

REFERENCES

1. Bureau d'Enquêtes et d'Analyses (BEA), Final Report on the accident on 1 June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF447 (July 2012) — full report: flight, aircraft, and the known pitot-icing vulnerability with retrofit pending.
2. BEA, Final Report AF447 (2012) — accident sequence: autopilot/autothrust disconnection, degraded control law, sustained nose-up input, stall, and the stall-warning logic dropping out at high angle of attack.
3. BEA, Final Report AF447 (2012) — probable cause: airspeed inconsistency from pitot icing, inappropriate control inputs, and failure to recognize and recover from the stall.
4. BEA, Final Report AF447 (2012) — finding on warning-system ergonomics and stall-training conditions (quoted).
5. G. Palmer / E. Strickland, "Air France Flight 447 Crash Caused by a Combination of Factors," IEEE Spectrum (2014) — analysis of the divergence between the trained and operational envelopes.
6. BEA, Final Report AF447 (2012), safety recommendations — pitot-probe replacement, manual high-altitude flying, approach-to-stall and stall recovery, unreliable-airspeed procedures, and angle-of-attack indication.
7. European Union Aviation Safety Agency, Upset Prevention and Recovery Training (UPRT) requirements for airline pilots (phased in by 2019); ICAO, Manual on Aeroplane Upset Prevention and Recovery Training (Doc 10011).
8. Federal Aviation Administration, FAA-S-ACS-1, Airline Transport Pilot Practical Test Standards and the 2014 Part 121 stall-recovery training rule reflecting BEA recommendations; FAA Advisory Circular 120-109A, Stall Prevention and Recovery Training.

THE LEARNING ENGINEERING LENS

LE INSIGHT

AF447 is the canonical case of training that matched one envelope of operation perfectly and another not at all. Stall **prevention** training was excellent. Stall **recovery from cruise altitude** training did not exist because the simulators of the era could not faithfully produce it. Layered on top was an automation-to-human handoff that arrived in a degraded control law the crew had never flown and a sidestick architecture that hid one pilot's inputs from the other. The pilots performed exactly as trained. The training was the wrong training, and the reform had to reach the regulator, because no single airline could close the gap unilaterally.

LENS APPROACH

AF447 is the worked example of induced sub-competency 1.2 (capability envelope at the edge of training) and the LENS D2/PT4 pairing — Iterative Development applied to training-program updates that must keep pace with operational regimes. Students use the case to specify training requirements from degraded-mode operational analysis (LENS D1), to design the evidence the BEA used to identify a training-design rather than airmanship failure (LENS D3), and to examine the human-AI handoff as a capability problem (LENS D5): the autopilot disconnection was a transition crews were trained to avoid rather than to handle. The case pairs with Kegworth (Case 103) as the canonical pair on transition-training failure across a changed system, and with the Boeing 737 MAX MCAS sequence as the near-current echo. LEO mapping: LEO-2

(Iterative Development) primary; LEO-4 (Test and Evaluation) for the BEA investigation framing; LEO-3 (Human-System Collaboration) for the automation-handoff dimension.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
AVIATION →	T H →	D1 **D2** D3 D4 D5	4	1.2	2 · 4
		Iterative Development			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Define the operational capability envelope explicitly — including degraded-automation and high-altitude regimes — and train to its edges, not only to the routine center.
- Engineer warning logic so that cues never punish the correct recovery input; validate stall-warning behavior across the full angle-of-attack range before fielding.
- Treat manual-flying proficiency in degraded modes as a measured deliverable, not an assumed residual of automated operation.

IN OPERATION / AFTER

- Audit recurrent training against the actual operational record so the trained envelope cannot silently diverge from the regimes crews encounter.
- Monitor for skill atrophy where reliable automation reduces hands-on exposure, and refresh manual competence before it erodes.
- Sustain the reform at the regulatory level (mandatory upset recovery) so a fix proven in one carrier propagates to all rather than lapsing.

REFLECTION QUESTIONS

1. The simulators of 2009 could not produce high-altitude stall behavior. What is the equivalent gap in your domain — the operational regime your training environment cannot reproduce?
2. Design the recurrent-training curriculum that would have caught the AF447 gap. Be specific about cost, evidence, and what makes the curriculum falsifiable against the operational record.
3. The autopilot handed control to a crew at the one moment it was least prepared to take it. Identify an automation-to-human handoff in your domain that occurs precisely when the human is least ready, and design the trigger or warning that would change that.

WHO BUILDS THIS learning science & instructional design · human factors & interaction design — plus domain experts and a learning engineer to integrate. TOOLS scenario simulation · competency models · usability testing · cognitive walkthroughs

CASE 112 AVIATION 1972 – PRESENT

NASA Saturn V Documentation

K Knowledge & Institutional Memory

IMPACT **Foundational U.S. aerospace case for the cost of losing the institutional capability to build a system**

IN BRIEF

When NASA returned to heavy-lift rockets in the 2000s, it confronted an uncomfortable fact: the institutional capability to build a Saturn V — the rocket that sent Apollo to the Moon — had been lost. The drawings and documents survived; the practical knowledge of how the components were manufactured, tested, and assembled, and why each choice was made, had walked out with the workforce that retired by the 1990s. The vehicle could be redesigned but not reproduced. Apollo was documented to an unprecedented degree, yet the documentation was not the capability: when the engineers who built the system left, the system left with them. Saturn V is the book's strongest evidence that institutional capability lives in people and sustaining practices, not in the artifacts an institution leaves behind.

BACKGROUND

The Saturn V that launched Apollo to the Moon was, by any measure, one of the best-documented engineering programs in history — exhaustive drawings, specifications, and test records.[0] That thoroughness is what makes the case bite: the program left behind close to everything a successor could ask for on paper, so any later difficulty in rebuilding it cannot be blamed on a thin archive but must lie in what the archive could not hold.

WHAT HAPPENED

Yet by the mid-2000s, as NASA worked the Constellation program, it had become apparent that the institutional capability to build a Saturn V had been lost. The drawings existed; the practical knowledge of how the components were manufactured, tested, and assembled — and why particular choices had been made — had walked out with the workforce that retired by the 1990s. The vehicle could be redesigned. It could not be reproduced.[1] The distinction is exact: redesign starts from requirements and rebuilds the reasoning afresh, while reproduction needs the original making-knowledge, and it was that knowledge — not the paperwork describing it — that had left with the people who held it.

THE INVESTIGATION

The case is canonical for the difference between documentation and institutional capability. Apollo's documentation was unprecedented, but it captured the **what**, not the tacit **how** — the judgment, the workarounds, the undocumented reasons — that lived in the people who did the work. When they left, that knowledge was in no archive to recover.[2] A drawing records the decision but not the deliberation behind it; the workaround that made a part

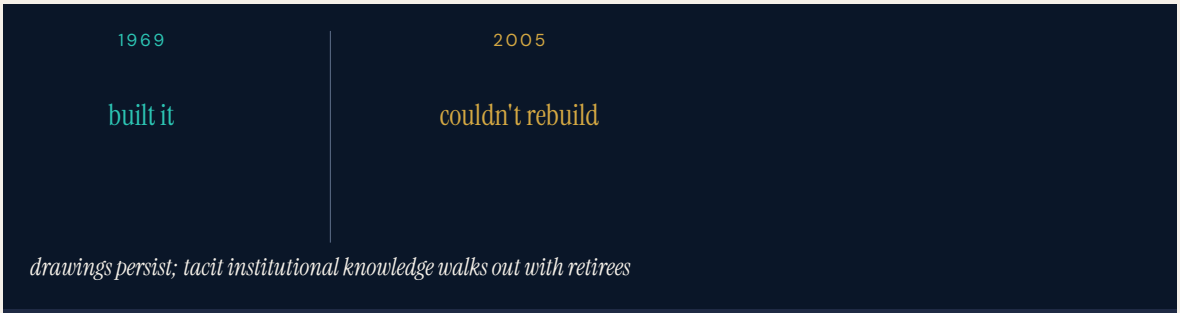
manufacturable and the reason a tolerance was set where it was lived in the doing, and once the doers retired there was no document from which to reconstruct it.

THE CAPABILITY GAP

Saturn V is the strongest available evidence that institutional capability is not the same as the artifacts an institution produces. Capability lives in people and in the practices that sustain them; a library of drawings is a necessary record but not a transferable ability. An institution that treats documentation as preservation of capability is, quietly, letting the capability expire.[3] The danger is that the archive looks like insurance: its very completeness can reassure managers that the capability is safe, so the people and practices that actually carry it are allowed to disperse precisely because the paperwork seems to stand in for them.

AFTERMATH & REFORM

The lesson reshaped how serious programs think about knowledge retention — apprenticeship, continuity of teams, deliberate capture of tacit reasoning, and not letting a critical capability rest in a handful of soon-to-retire heads.[4] It pairs with the chapter's other memory cases: knowledge, unlike a document, has to be actively kept alive, or it is gone in a generation. Each of the retention practices addresses the same root cause the Saturn V exposed — that tacit making- knowledge transfers only person to person — so apprenticeship and team continuity are not nice-to-haves but the actual mechanism by which a capability outlives the people who first held it.



SATURN V — DOCUMENTATION IS NOT CAPABILITY

> *The Saturn V drawings exist. The Saturn V does not.*
>
> PARAPHRASING THE NASA CONSTELLATION-ERA CAPABILITY DISCUSSION, C. 2005

REFERENCES

1. NASA Constellation Program documentation and reviews (2005–2010) — the difficulty of reproducing Saturn V capability.
2. R. E. Bilstein, Stages to Saturn: A Technological History of the Apollo/Saturn Launch Vehicles (NASA SP-4206, 1980) — the program and its workforce.
3. The documentation-vs-capability distinction (editors' synthesis of the Saturn V record).
4. M. Polanyi, The Tacit Dimension (1966); I. Nonaka & H. Takeuchi, The Knowledge-Creating Company (1995).
5. J. S. Brown & P. Duguid, The Social Life of Information (2000) — tacit and institutional knowledge.

The Saturn V case is the canonical evidence that institutional knowledge is not equivalent to documentation. The documents persist; the capability does not. Capability engineering must treat the people who hold tacit knowledge as a primary institutional asset.

LENS APPROACH

LENS uses the Saturn V case in LEN 8 to teach that retaining and overlapping the expert personnel who hold tacit making-knowledge is itself an engineered deliverable (induced 6.3) — not a by-product of writing better documents. The capability to rebuild a Saturn V was lost because workforce overlap and retention were never treated as a designed retention deliverable, and documentation alone cannot carry tacit process knowledge across a generation. The teaching point in LEN 1 is therefore not "document more thoroughly" but "engineer the personnel continuity that the archive can never substitute for."

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
AVIATION →	K →	D1 D2 D3 D4 **D5**	2	6.3	2

Navigating Sociotechnical Constraints

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Capture tacit reasoning as the work happens — the why behind a choice, the workaround that made a part buildable — not just the resulting drawing.
- Build capability into teams and apprenticeship, so the making-knowledge has a living carrier rather than resting in a handful of soon-to-retire heads.
- Treat documentation as a record, never as a substitute for the people and practices that hold the capability.

IN OPERATION / AFTER

- Audit critical capabilities for single points of human failure — knowledge held by a few near-retirement practitioners — and transfer it before they leave.
- Periodically test reproducibility, not just whether the archive is complete, since a full archive can mask a capability that has already expired.
- Sustain continuity of teams and practice deliberately, so a capability is kept alive across generations rather than rediscovered at need.

REFLECTION QUESTIONS

1. Identify a capability in your domain that is currently held in the tacit knowledge of a small number of senior practitioners. What is your institution's capacity to reproduce it after they retire?

2. Design the institutional practice that would preserve a capability against the retirement of its holders.

3. Saturn V was exhaustively documented yet could not be reproduced, because the archive captured the **what** and not the tacit **how**. Identify a capability in your institution whose documentation might be giving false reassurance — and describe what the paperwork is failing to capture.

WHO BUILDS THIS knowledge management & data engineering — plus domain experts and a learning engineer to integrate. TOOLS knowledge bases · data pipelines

Chapter 6

Aviation & Aerospace — What Works — and the Frontier

When an industry engineers its own error rate down and keeps the books open.

Flying got safer when reporting became cheaper than concealment.

AN OBSERVATION RECURRING ACROSS THE CHAPTER'S CASES.

CASE 117 AVIATION 1981 – PRESENT

Crew Resource Management & CAST

[T] Training Gap [H] Human-System Interface [N] Normalization of Deviance

IMPACT 83% reduction in U.S. commercial aviation fatality risk (1998–2008); 95% reduction in fatalities per 100M passengers over 20 years

IN BRIEF

In March 1977, two 747s collided in fog at Tenerife and 583 people died — in part because a KLM flight engineer's twice-voiced doubt about the runway was overridden by his captain. The failure was not of skill but of the system by which skill in one seat reached another. Crew Resource Management, formalized by United Airlines in 1981, re-engineered the cockpit as a coordinated team: explicit communication protocols, named authority gradients, structured briefings. Twenty years later the Commercial Aviation Safety Team (CAST) added closed-loop analysis of operational data to find and fix hazards before they caused accidents. The pairing — cultural redesign plus continuous evidence — helped drive the 83% reduction in U.S. commercial-aviation fatality risk between 1998 and 2008 that CAST achieved alongside new aircraft and regulation, work that earned CAST the 2008 Collier Trophy.

BACKGROUND

By the 1970s, accident investigations were repeatedly finding that crashes stemmed not from a lack of individual flying skill but from breakdowns in how a crew worked together. The 1977 Tenerife disaster — 583 dead — was the starkest example: a KLM flight engineer questioned, twice and indirectly, whether the runway was clear, and the senior captain dismissed him and continued the takeoff. The pattern that emerged was consistent across investigations — skilled crews failing not because anyone lacked competence, but because the crew's most senior voice could close off information held by a more junior one before it reached the decision.[0]

THE INTERVENTION

Crew Resource Management, formalized by United Airlines in 1981, treated crew coordination as an engineerable property of the system rather than a matter of personality. It introduced explicit communication protocols, named and trained against authority gradients, and instituted structured briefings and debriefings — rebuilding the cockpit as a team in which information from any seat could reach the decision. Rather than exhorting captains to listen better, it built coordination into the standard procedure itself, so the behavior that Tenerife had lacked became the trained default rather than a matter of individual temperament.[1]

HOW IT WORKED

CRM did not teach individual airmanship; it engineered the system of interaction in which airmanship operates. By making it legitimate and expected for a junior officer to challenge a captain, and by giving crews a shared protocol for doing so, it closed the path by which

Tenerife-style deference had absorbed safety-critical information instead of transmitting it. The change was structural rather than attitudinal: a challenge that had once depended on a junior officer's nerve now had a named procedure behind it, so the same doubt that went unheard at Tenerife had an authorized route to the decision.[2]

THE EVIDENCE

Twenty years on, the Commercial Aviation Safety Team added the missing second half: closed-loop hazard identification on operational data, prioritized enhancements, tracked implementation, and measured outcomes. CAST set an 80% fatality-reduction target and exceeded it, reaching 83% by 2007 — work recognized with the 2008 Collier Trophy. Over twenty years, fatalities per 100 million passengers fell roughly 95%. The loop closed on itself: data surfaced the next hazard, enhancements were prioritized against it, and the measured outcome fed back into the priorities, so improvement continued rather than plateauing once the cultural change had taken hold.[3]

WHAT TRANSFERRED

CRM and CAST together define what a mature capability-engineering apparatus looks like: a cultural redesign paired with a continuous-evidence loop, where neither half works alone. The model has been exported to surgery, firefighting, and other high-consequence domains, and is now the template for redesigning human roles in AI-augmented systems. What transferred was not the specific protocols but the design logic itself — that coordination is an engineerable system property and that the cultural change needs a measurement loop to keep it honest over time.[4]

AUTHORITY GRADIENT FLATTENED; COMMUNICATION CLOSES THE LOOP

> *CRM succeeded because it treated crew coordination as an engineerable property of the system.*
>
> PARAPHRASING HELMREICH & FOUSHEE, IN KANKI ET AL., CREW RESOURCE MANAGEMENT (2010)

REFERENCES

1. FAA Advisory Circular 120-51E, Crew Resource Management Training — CRM protocols and authority-gradient training.
2. Helmreich, Merritt & Wilhelm (1999), "The Evolution of Crew Resource Management Training in Commercial Aviation," International Journal of Aviation Psychology.
3. Spanish CIAIAC / ALPA reports on the 1977 Tenerife collision — the overridden crew challenge.
4. CAST/FAA Safety Enhancement reports (2016, 2018) — the closed-loop data process and the 83% reduction.
5. Collier Trophy citation (2008); Kanki, Helmreich & Anca (2010), Crew Resource Management.

THE LEARNING ENGINEERING LENS

LE INSIGHT

CRM is the canonical evidence that capability is engineerable at the system level, not just the individual. Tenerife was not solvable by hiring better pilots — only by changing the authority structure inside the cockpit. The intervention worked because it was **paired**: a procedural change plus a cultural change in how the procedure was authorized. Either alone fails.

LENS APPROACH

LENS treats CRM/CAST as the foundational success case across the curriculum. LEN 1 uses it as the problem-framing exemplar; LEN 4 uses CAST as the model closed-loop evidence system; LEN 2 uses CRM as the template for redesigning human roles in automated environments — the logic now being applied to AI-augmented systems.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
AVIATION →	T H N →	D1 D2 **D3** D4 D5	3	4.3	3
		Human-System Collaboration			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Build the coordination behavior into standard procedure — explicit communication protocols, named authority gradients, and structured briefings — so it is the trained default, not a matter of temperament.
- Authorize the junior-to-senior challenge explicitly and drill it, so safety-critical information has a procedural route to the decision rather than depending on individual nerve.
- Stand up the evidence half from the start: instrument operational data so hazards can be found and prioritized before they cause an accident.

IN OPERATION / AFTER

- Run the closed loop continuously — surface hazards from data, prioritize enhancements, track implementation, and measure outcomes — so improvement does not plateau once the culture shifts.
- Set and exceed an explicit reduction target (the CAST model) so the intervention is judged against a measured outcome, not against its own activity.
- Export the design logic, not just the protocols, when transferring to new domains, adapting the paired cultural-plus-evidence structure to each setting.

REFLECTION QUESTIONS

1. Identify a high-consequence domain whose authority gradient absorbs information rather than transmitting it. What would the CRM equivalent intervention look like there?
2. CRM is paired with CAST. What is the closed-loop evidence system in your domain — and if there is not one, what would it cost to build?
3. CRM made a junior officer's challenge a named procedure rather than an act of nerve. Identify a moment in your domain where the right information depends on someone's courage, and design the protocol that would make raising it the expected default instead.

WHO BUILDS THIS learning science & instructional design · human factors & interaction design · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS scenario simulation · competency models · usability testing · cognitive walkthroughs · incident analysis · safety dashboards

CASE 119 AVIATION 1976 – PRESENT

Aviation Safety Reporting System (ASRS)

T Training Gap K Knowledge & Institutional Memory N Normalization of Deviance

IMPACT NASA-administered confidential reporting system; more than 2M reports received; foundational architecture for evidence-driven aviation safety

IN BRIEF

The Aviation Safety Reporting System, run by NASA on behalf of the FAA since 1976, accepts confidential reports from pilots, controllers, mechanics, and cabin crew about incidents, near-misses, and safety concerns. Its load-bearing feature is institutional, not technical: reporting an event to ASRS confers immunity from FAA enforcement for the conduct reported, within specified limits, making honest reporting the rational choice. Over nearly fifty years and more than two million reports, ASRS has become the world's largest repository of aviation-safety information, surfacing patterns — automation surprise, runway incursions, fatigue effects — before they reached formal investigation thresholds. The architecture has been emulated across domains, and is the canonical success case for an evidence system paired with a credible commitment to non-punitive use.

BACKGROUND

The most valuable safety information in any high-consequence domain lives with front-line operators — the near-misses and quiet errors that never reach an accident report. But operators will not surrender that information to an institution that can punish them for it, so the data that could prevent the next accident stays locked in the people who hold it. The incentives run backward: the person best placed to report a near-miss is the same person a punitive system gives the strongest reason to stay silent, so the data that matters most is the data least likely to surface.[0]

THE INTERVENTION

In 1976 the FAA and NASA created the Aviation Safety Reporting System, a confidential channel for pilots, controllers, mechanics, and cabin crew to report incidents, near-misses, and concerns. NASA — not the regulator — administers it, and reporting an event confers immunity from FAA enforcement for the conduct reported, subject to specified limits. Putting a neutral party between the reporter and the enforcer directly addressed the backward incentive — an operator now had a positive reason to report, because doing so converted potential jeopardy into protection.[1]

HOW IT WORKED

The system pairs a technical artifact (the reporting form and a searchable database) with a cultural commitment (protected, non-punitive use). The immunity provision makes reporting the rational choice for the operator, and NASA's role as a neutral third party makes the protection credible. Either half alone fails; together they make the data flow to the institution

that can act on it. The credibility of the protection is what does the work — a promise of non-punishment from the regulator itself would be doubted, so routing it through NASA is what makes operators trust it enough to report.[2]

THE EVIDENCE

Over nearly fifty years ASRS has accumulated more than two million reports — the largest single repository of aviation-safety information in the world. Patterns such as automation surprise, runway incursions, and fatigue effects were first identified at scale through ASRS data before they crossed formal investigation thresholds. Surfacing a pattern before it reaches an accident is the whole point — the value of the system is the hazards it lets the industry act on while they are still near-misses, not the reports it collects after the fact.[3]

WHAT TRANSFERRED

ASRS has been studied and emulated across domains — patient-safety reporting systems, the maritime and aviation CHIRP scheme, and similar systems in rail and nuclear power. It is the canonical positive case for evidence architecture paired with an institutional commitment to non-punitive learning, the defining design pattern of a "just culture." The breadth of emulation underscores that the load-bearing element travels — wherever the most valuable safety data sits with operators who fear punishment, the same protected-reporting design recurs as the way to unlock it.[4]

ASRS — PAIRED TECHNICAL ARTIFACT AND PROTECTED REPORTER

> *ASRS is the model for confidential, voluntary, non-punitive incident reporting in any high-consequence domain.*
>
> PARAPHRASING JAMES REASON, MANAGING THE RISKS OF ORGANIZATIONAL ACCIDENTS, 1997

REFERENCES

1. NASA ASRS Program documentation and annual reports — system design, immunity provision, and report volume.
2. Reason, J. (1997), Managing the Risks of Organizational Accidents — non-punitive reporting as a model (paraphrased).
3. NASA ASRS technical reports (Connell et al.) — patterns first surfaced via ASRS data.
4. Dekker, S. (2012), Just Culture — the cultural commitment to non-punitive use.
5. CHIRP and patient-safety reporting-system documentation — cross-domain emulation.

THE LEARNING ENGINEERING LENS

LE INSIGHT

ASRS is the canonical positive case for paired-intervention evidence architecture. The technical artifact (the reporting form and the database) is paired with the cultural commitment to non-punitive use. Either alone fails. Together they have produced the most comprehensive operational-safety dataset in any domain.

LENS APPROACH

LENS uses ASRS in LEN 4 as the foundational positive case for evidence architecture and in LEN 8 for institutional commitment to non-punitive learning. Studio projects design ASRS-equivalents for new domains.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
AVIATION	T K N	D1 D2 D3 **D4** D5	2	4.2	4
		Test and Evaluation			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Pair a simple reporting artifact (a form and searchable database) with a credible commitment to non-punitive use, since the channel without the protection collects nothing of value.
- Confer immunity for reported conduct so that reporting becomes the rational choice, directly reversing the incentive that otherwise keeps the most valuable data hidden.
- Route the protection through a neutral third party rather than the enforcer, so operators trust the non-punitive promise enough to report against their own interest.

IN OPERATION / AFTER

- Analyze the accumulated reports to surface patterns — automation surprise, runway incursions, fatigue — and act on them while they are still near-misses rather than accidents.
- Protect the immunity provision over time, since a single high-profile punishment of a reporter would collapse the trust the whole system depends on.
- Export the protected-reporting design, not just the database, to new domains, adapting the neutral-party structure wherever valuable safety data sits with operators who fear punishment.

REFLECTION QUESTIONS

1. Identify a domain in your institution that would benefit from an ASRS-equivalent. What cultural commitment would be required for it to function?
2. Design the institutional commitment that makes an ASRS-equivalent operational rather than merely declared.
3. ASRS made the protection credible by routing it through NASA rather than the regulator. Identify a reporting channel in your domain that operators distrust, and specify the neutral party or structural separation that would make its non-punitive promise believable.

WHO BUILDS THIS learning science & instructional design · knowledge management & data engineering · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS scenario simulation · competency models · knowledge bases · data pipelines · incident analysis · safety dashboards

CASE 120 AVIATION SAFETY ENGINEERING HUMAN FACTORS 1996 – 2002

EGPWS / TAWS — Closing the CFIT Category in Commercial Aviation

H Human-System Interface K Knowledge & Institutional Memory G Governance & Trust

IMPACT Honeywell's Enhanced Ground Proximity Warning System (EGPWS, 1996), mandated by the FAA as Terrain Awareness and Warning System (TAWS) for US-registered turbine aircraft beginning in 2000 and broadly worldwide by 2002, converted controlled flight into terrain (CFIT) — historically one of the largest categories of commercial-aviation fatalities — into a category whose rate in equipped fleets has fallen sharply; CFIT events on properly equipped and operating airliners are now rare

IN BRIEF

Controlled flight into terrain (CFIT) — a serviceable aircraft under the pilot's control flown unintentionally into the ground, water, or an obstacle — was for decades one of the largest categories of commercial-aviation fatalities. The 1979 Air New Zealand Mt Erebus crash (257 dead) and the 1995 American Airlines 965 Cali crash (159 dead) are canonical examples. The first-generation Ground Proximity Warning System (GPWS) developed by C. Donald Bateman at Sundstrand / Honeywell in the 1970s used radio altimeter and rate-of-descent inputs to warn of imminent terrain contact; it reduced CFIT but produced late warnings and was blind to terrain ahead of the aircraft. Enhanced GPWS (EGPWS), introduced commercially in 1996, added a digital terrain database and aircraft position to the input set, permitting forward-looking terrain-avoidance alerting. The FAA mandated EGPWS-class equipment (formally TAWS) on US-registered turbine aircraft beginning March 2000, with full compliance required by 2005; ICAO and most national regulators followed. The published outcome record is that CFIT in EGPWS-equipped commercial fleets has become rare — NTSB, FAA, and Flight Safety Foundation analyses consistently report a sharp decline. The hedge that survives: residual CFIT events still occur, typically involving disabled or inhibited equipment, deviation from procedure, or terrain outside the database, and the case has to honor the system- in-operation discipline rather than the system-as-installed claim.

BACKGROUND

Through the 1960s and 70s, controlled flight into terrain was one of the highest-fatality categories in commercial aviation. The pattern was structurally consistent: serviceable aircraft, qualified crew, often in IMC (instrument meteorological conditions) or at night, flown into rising terrain the crew had not visualized correctly. The Air New Zealand Mt Erebus crash (1979, 257 dead) and the American Airlines 965 Cali crash (1995, 159 dead) are the canonical

examples — competent crews who lost situational awareness about their position relative to terrain in conditions that prevented visual recovery.[0]

THE INTERVENTION

The first engineered intervention was the Ground Proximity Warning System (GPWS), developed in the early 1970s by C. Donald Bateman at Sundstrand (later Honeywell). GPWS used radio altimeter readings and rate-of-descent to generate "pull up" and similar warnings when the aircraft was descending toward terrain directly below it. GPWS reduced CFIT meaningfully through the 1970s and 80s, but had two structural limits: it produced late warnings (the aircraft was already close to terrain when the alert fired), and it was blind to terrain ahead of the aircraft — the Cali accident occurred in a GPWS-equipped aircraft because the rising terrain was ahead of the flight path, not below.[1]

HOW IT WORKED

Enhanced GPWS (EGPWS), introduced by Honeywell in 1996, addressed both limits by adding a digital terrain database and aircraft position (GPS / IRS) to the input set. The system can now compute a forward-looking terrain surface relative to the aircraft's projected flight path and provide alerts well before terrain contact is imminent. The FAA codified the capability in the Terrain Awareness and Warning System (TAWS) regulation, requiring TAWS-class equipment on US-registered turbine aircraft with six or more passenger seats beginning March 29, 2000, with full equipage by 2005. ICAO and most national regulators followed with parallel mandates.[2]

THE EVIDENCE

The published outcome record across NTSB accident statistics, FAA reporting, and Flight Safety Foundation analyses is that CFIT in EGPWS-equipped commercial fleets has fallen sharply. The category that once dominated airliner-fatality statistics is now an uncommon-event category in equipped fleets. The structural claim the case makes is the cue/alert-design one: a failure mode in which the operator's perception of terrain was the limiting variable was converted into a monitored, recoverable mode by surfacing the forward terrain picture as an actionable alert. Pair with anesthesia monitoring (Case 27) at the cue/alert-as- capability layer, and with TCAS (Case 121) at the automated-advisory-system layer.[3]

WHAT TRANSFERRED

The hedge has to survive into the case. CFIT events still occur, typically involving one or more of: EGPWS disabled or inhibited (crew action, MEL release, maintenance), deviation from procedure (e.g., descent below minimum sector altitude under pressure), or terrain or obstacles not represented in the database (rapidly changing wind-turbine and structure environments are a known frontier). The system-in- operation has to be flying and the crew has to act on the alert; a rule-of-thumb in the safety community is that EGPWS is most useful when its warnings are taken seriously enough that they are rare in operation. The case teaches the cue/alert intervention at its most durable, with the qualification that the capability depends on the standard being honored in operation, not on the equipment being installed.[4]

> *The capability depends on the standard being honored in operation, not on the equipment being installed.*

EDITORS' SYNTHESIS OF FAA TAWS RULE HISTORY AND FSF ALAR ANALYSES.

REFERENCES

1. Bateman, C. D. (1999), "The Introduction of Enhanced Ground Proximity Warning Systems (EGPWS) into Civil Aviation Operations Around the World," in Proceedings of the 11th Annual European Aviation Safety Seminar, Flight Safety Foundation, pp. 259–274 — developer history.
2. Federal Aviation Administration (2000), 14 CFR §§ 91.223, 121.354, 135.154 — Terrain Awareness and Warning System (TAWS) equipage requirement.
3. Flight Safety Foundation (1998 – 2000), CFIT / ALAR Task Force reports — operational and outcome analyses motivating mandate.
4. NTSB (1996), Aircraft Accident Report AAR-96-05, American Airlines 965, Cali, Colombia, December 20 1995.
5. Royal Commission to Inquire into the Crash on Mount Erebus, Antarctica of a DC10 Aircraft Operated by Air New Zealand Limited (1981), final report (Mahon report).

THE LEARNING ENGINEERING LENS

LE INSIGHT

EGPWS / TAWS is the canonical cue/alert intervention at fleet scale. The forward-looking terrain alert converted a failure mode in which the crew's terrain perception was the limiting variable into a monitored, recoverable mode. CFIT in equipped fleets has become rare; residual events typically involve inhibited equipment or procedure deviation, and the hedge is the case.

LENS APPROACH

EGPWS / TAWS is the aviation cue/alert intervention case (induced 3.1; LENS D4/PT5) — Domain 4 for cue-design-as- deliverable; Domain 3 for the operator-cue boundary. Pair with TCAS (Case 121) and Case 27 (anesthesia monitoring).

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
AVIATION SAFETY ENGINEERING HUMAN FACTORS	H K G	D1 D2 D3 **D4** D5 Test and Evaluation	5	3.1	4 · 3

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Identify the operator-perception variable that is the limiting variable in a failure mode (here: the crew's awareness of terrain ahead of the aircraft) and engineer the system that surfaces that variable as an actionable alert with enough lead time to recover.
- Pair the alert design with a regulatory mandate that makes the equipment non-waiverable across the fleet, so adoption is fleet-level capability rather than per-operator choice. The 1996-to-2001 gap between commercial availability and mandate is the political-process cost.
- Build the cue's lead time around the time the operator needs to act, not the time the equipment can produce the alert; a too-late alert is the GPWS limitation EGPWS was specifically built to address.

IN OPERATION / AFTER

- Carry the system-in-operation hedge into communication: the capability depends on EGPWS being operational, not inhibited, and on the crew acting on the alert. Inhibition discipline is part of the deliverable.
- Maintain the terrain database as a continuously updated artifact; the equipment as installed is only as good as the database it queries, and rapidly changing obstacle environments (wind turbines, structures) are a known frontier.
- Treat residual CFIT events as evidence about the operational discipline, not as evidence against the intervention; the system that has reduced a fatality category to rare uncommon events is doing the work the case claims.

REFLECTION QUESTIONS

1. Identify a failure mode in your domain where operator perception of an external variable is the limiting factor. What is the analog of the digital terrain database — the engineered representation of the variable — and what lead time would the cue need to be actionable?

2. Specify the regulatory or institutional mandate path you would expect: EGPWS reached the market in 1996, was mandated in 2001, and was fully equipaged by 2005. Five years from commercial availability to full equipage is a useful planning datum for a fleet-scale capability mandate.

3. The system-in-operation hedge is binding. What inhibition discipline would your program require so that the engineered recovery layer is operating when the failure mode appears, and how would you instrument that the discipline is being honored?

WHO BUILDS THIS human factors & interaction design · knowledge management & data engineering · governance, ethics & measurement — plus domain experts and a learning engineer to integrate. TOOLS usability testing · cognitive walkthroughs · knowledge bases · data pipelines · governance frameworks · bias & evidence audits

Chapter 7

Defense & National Security — What Fails

When the capability the mission requires goes unspecified at the point of contact.

The system was fielded. The operator's task was not.

AN OBSERVATION RECURRING ACROSS THE CHAPTER'S CASES.

CASE 124 DEFENSE 2017

USS Fitzgerald & USS John S. McCain

T Training Gap K Knowledge & Institutional Memory N Normalization of Deviance

IMPACT 17 sailors killed in two avoidable destroyer collisions in the Western Pacific within nine weeks

IN BRIEF

In the summer of 2017 two U.S. Navy destroyers of the forward-deployed Seventh Fleet collided with merchant ships nine weeks apart, killing seventeen sailors — seven on the Fitzgerald, ten on the John S. McCain. Their crews' seamanship, navigation, and console skills had eroded under a decade of relentless operational tempo and the self-study "training" that replaced the in-person school the Navy cut in 2003. Investigations by the Navy and the NTSB judged both collisions avoidable and traced them to insufficient training and weak oversight; on the McCain, a confusing touch-screen helm let a watch team believe it had lost steering it never actually lost. The case is this book's canonical training gap: stated readiness and real readiness diverged for years, and the system could no longer see the difference.

BACKGROUND

The destroyers belonged to the forward-deployed Seventh Fleet in Japan, kept at the fleet's highest tempo — where training was the first thing spent. Unlike home-ported ships, which rotate through a dedicated work-up cycle before deploying, the forward-deployed force was expected to be continuously available, so there was rarely a quiet stretch in which to recover a lapsed qualification. In 2003 the Navy replaced its in-person Surface Warfare Officers School course with self-study CD-ROMs — "SWOS in a Box" — sending newly commissioned officers directly to their ships to learn the trade as time and the ship's qualified watchstanders allowed.[0] The change was framed as a cost-saving modernization; in practice it transferred junior-officer education from the schoolhouse to the bridge at the moment the operational tempo on those bridges left the least slack to absorb it. By 2017 the GAO found 37% of the Japan-based ships' warfare certifications — including basic seamanship — expired, a fivefold rise since 2015, with the lapses routinely waived rather than fixed.[1]

WHAT HAPPENED

On 17 June 2017 the Fitzgerald, the give-way vessel in a busy shipping channel, took no early action and was struck by the container ship ACX Crystal off the coast of Japan; the sea poured into a berthing compartment where the crew slept, and seven sailors drowned before they could escape.[2] The NTSB later described a bridge team that had lost track of converging traffic on a clear night, and an officer of the deck whose qualifications and recency on the very procedures the situation demanded had lapsed under the waiver regime. Nine weeks later the John S. McCain collided with the tanker Alnic MC near Singapore, killing

ten: while shifting throttle control between consoles, a watchstander unknowingly handed off steering to another station, the ship turned across the strait's traffic, and no one on the bridge understood the touch-screen helm well enough to recognize what had happened.[3] For more than a minute the bridge team believed the ship had lost steering it had never lost — a misdiagnosis the interface invited and the training had not equipped anyone to overturn.

THE INVESTIGATION

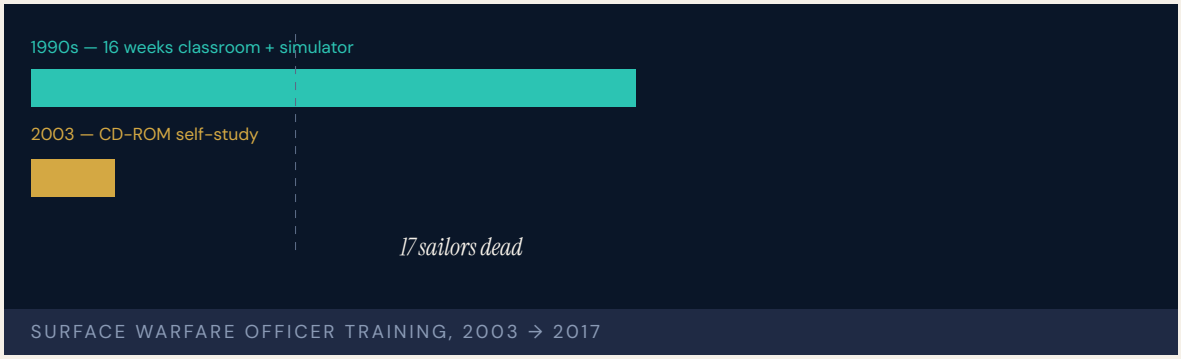
The Navy's Comprehensive Review (2017) judged both collisions avoidable, citing failures in basic seamanship, navigation, and operating the ships' own equipment.[4] The NTSB found the McCain's probable cause to be a lack of Navy oversight that produced insufficient training and inadequate bridge procedures,[5] and faulted the design of the touch-screen helm, installed to modernize the bridge, whose blending of steering and throttle made an unintended transfer of control easy to trigger and hard to notice — a trap waiting for an undertrained crew. The companion Strategic Readiness Review, commissioned by the Secretary of the Navy and led by retired Admiral Gary Roughead and Michael Bayer, reached further into the institution: a decade of "can-do" responses to mounting demand had eroded the manning, certification, and maintenance margins the surface force was built on, and the readiness reports senior leaders relied on had stopped reflecting the conditions on the ships.[6] Watchbills and crew-day logs gathered after the collisions showed watchstanders routinely averaging fewer than five hours of sleep on patrol — a finding the NTSB folded into its causal chain.

THE CAPABILITY GAP

The gap was invisible from inside. The Strategic Readiness Review described risks that "accumulated over time and did so insidiously," the system no longer able to see that the processes meant to surface shortfalls had themselves failed.[7] Each individual waiver was locally defensible — a deadline met, a deployment kept — but in aggregate they hollowed out the force while every readiness dashboard still reported green. Stated and actual readiness had diverged for a decade; the cost arrived as seventeen lives, not a failed inspection. Two collisions, nine weeks apart, ruled out the comforting reading that one was an outlier: Fitzgerald taught that an under-trained bridge team could fail at the most basic give-way rules, and McCain taught that an unfamiliar interface could be the precipitating mechanism through which the same gap reached the hull. The pair, not either case alone, made the diagnosis structural rather than individual.

AFTERMATH & REFORM

The reforms were the deepest in a generation: the in-person officer pipeline was rebuilt as a multi-phase Basic Division Officer Course, reinstating classroom and simulator instruction the 2003 CD-ROM model had displaced; a Ready-for-Sea Assessment and a Naval Surface Group Western Pacific stood up to give forward-deployed units the independent certification cycle home-ported ships already had; circadian watchbills were adopted fleet-wide to fight fatigue; and the touch-screen helm was slated for replacement by a conventional wheel and throttle across the destroyer fleet.[8] Each change conceded that the training and the interface had been real requirements all along — and that trading them for tempo had only moved the cost onto two ships full of people.[9]



SURFACE WARFARE OFFICER TRAINING, 2003 → 2017

> *The risks that were taken in the Western Pacific accumulated over time and did so insidiously.*
>
> U.S. NAVY STRATEGIC READINESS REVIEW, 2017

REFERENCES

1. T. C. Miller, M. Faturechi & R. Rotella, "Years of Warnings, Then Death and Disaster," ProPublica (2019) — the 2003 SWOS-in-a-Box shift.
2. GAO, Navy Readiness, GAO-17-809T (Sept. 2017) — 37% of Japan-based warfare certifications expired by June 2017.
3. NTSB, Collision between USS Fitzgerald and MV ACX Crystal, NTSB/MAR-20/02 (2020), DCA17PM018.
4. NTSB, Collision between USS John S. McCain and Tanker Alnic MC, NTSB/MAR-19/01 (2019), DCA17PM029.
5. U.S. Navy, Comprehensive Review of Recent Surface Force Incidents (Nov. 2017) — both collisions judged avoidable.
6. NTSB/MAR-19/01 (2019) — probable cause: insufficient training and inadequate bridge procedures from a lack of Navy oversight.
7. R. Rotella et al., "The Navy Installed Touch-Screen Steering Systems to Save Money," ProPublica (2019).
8. U.S. Navy, Strategic Readiness Review (Dec. 2017) — risks that "accumulated over time and did so insidiously."
9. U.S. Navy corrective actions (2017); NTSB/MAR-19/01 recommendations; USNI News (2017–2020).
10. Surface Warfare Officers School Command, Basic Division Officer Course curriculum and the post-2017 return to in-person instruction; Naval Surface Group Western Pacific stand-up (2019) as the forward-readiness certification authority.
11. NTSB/MAR-19/01 (2019) — watchstander fatigue findings, including average sleep hours and the touch-screen helm misdiagnosis sequence.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Fitzgerald/McCain is the canonical Training Gap case because the gap was invisible from inside the system. Operational tempo and self-study "training" combined to produce a fleet whose stated readiness and actual readiness diverged for more than a decade. The two collisions nine weeks apart converted what could have been read as an outlier into a structural diagnosis: an under-trained bridge team failing the most basic give-way rules, and an unfamiliar interface that made the same gap reach the hull. Seven sailors died on Fitzgerald; ten on McCain. The cost of the divergence was paid in lives long after it could have been measured in dollars or inspections.

LENS APPROACH

Fitzgerald/McCain is the worked example of induced sub-competency 1.1 (engineered vs. stated requirements) and the LENS D1/PT3 pairing — Systems Analysis applied to the capability-engineering problem of underway watchstanding. Students reconstruct the capability requirements from operational analysis, then design the evidence architecture (LENS D3) that would have surfaced the gap before the collisions and the sociotechnical reforms (LENS D4) that would keep waivers from quietly hollowing the requirement out. The case pairs with Three Mile Island (Case 166) as the failure that engineered durable industry reform through INPO. LEO mapping: LEO-1 (Systems Analysis) primary, LEO-5 (Sociotechnical Constraints) for the waiver-and-reporting institutional dynamics.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
DEFENSE →	T K N →	D1 D2 D3 D4 D5	3	1.1	1 · 5
		Systems Analysis			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Specify watchstanding proficiency as a measured deliverable — define the competency, instrument it, and gate qualification on demonstrated skill rather than sea time.
- Keep human-factors review in the procurement loop: validate that any interface change (e.g., the bridge console) is tested against operator performance before it is fielded.
- Design the readiness signal to report demonstrated capability, not self-attested course completion.

IN OPERATION / AFTER

- Audit the gap between reported and actual readiness with independent assessment (the Ready-for-Sea model) — with authority to pull a unit offline.
- Protect training time against operational tempo so the capability cannot erode silently when schedules tighten.
- Track leading indicators — qualification currency, near-miss reports — so divergence is visible before it is paid for in lives.

REFLECTION QUESTIONS

1. The Navy replaced classroom and simulator instruction with CD-ROM self-study in 2003. What measurement would have detected the capability shortfall before 2017?

2. The Strategic Readiness Review found the readiness-reporting system had itself stopped working. Design a capability-evidence pipeline that would not normalize its own gaps.

3. Identify a capability in your organization that is certified by completion (a checked box) rather than demonstrated performance. What measure would convert it to evidence — and who would hold the authority to act on a red signal?

WHO BUILDS THIS learning science & instructional design · knowledge management & data engineering · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS scenario simulation · competency models · knowledge bases · data pipelines · incident analysis · safety dashboards

CASE 129 DEFENSE 1991

Patriot Missile / Dhahran

D Designed Out H Human-System Interface K Knowledge & Institutional Memory

IMPACT 28 U.S. soldiers killed in their barracks; roughly 100 wounded

IN BRIEF

On 25 February 1991 a Scud missile struck a U.S. barracks in Dhahran, Saudi Arabia, killing 28 soldiers and wounding about a hundred; the Patriot battery defending the area never engaged it. Designed for short engagements against Soviet aircraft in Europe, the Patriot tracked time in a register whose tiny rounding error grew with every hour of uptime. After about a hundred hours of continuous operation the Gulf War demanded, the radar's range gate was off by a third of a second — enough to look in the wrong place. Israeli operators had flagged the drift two weeks earlier, and a patch reached Dhahran the day after the strike. The capability to hold accuracy under sustained use was assumed away at design time, and no one carried that assumption forward when the mission changed.

BACKGROUND

The Patriot was built to defend Western Europe against Soviet aircraft in engagements of minutes — switched on, fired, switched off, then moved before it could be targeted. In the 1991 Gulf War it was doing something else: defending fixed sites in Saudi Arabia against ballistic missiles, in batteries left running continuously for more than a hundred hours because the threat came without warning and could not be scheduled. The mismatch between the original concept of operations and the new use was real, and invisible to the operators, because no one had told them the machine had been built around an assumption they were now violating.[0]

WHAT HAPPENED

The system tracked time in a 24-bit fixed-point register, and a tiny rounding error — invisible in any short engagement — accumulated with every hour of uptime. After about a hundred hours the timing was off by roughly a third of a second, which seems negligible until it is multiplied by a ballistic missile's speed: enough for the radar's range gate to look in the wrong place and reject the real track as noise. On 25 February 1991 an incoming Scud arrived where the Patriot was no longer searching, passed unengaged, and struck a barracks in Dhahran, killing twenty-eight soldiers of the 14th Quartermaster Detachment and wounding about a hundred.[1]

THE INVESTIGATION

The drift was not unknown: Israeli operators had flagged it two weeks earlier from their own sustained use, and engineers had a patch already in hand — which reached Dhahran the day after the strike, too late to matter.[2] The only field mitigation was an advisory to reboot after "very long" run times, never defining "very long," so a crew could obey the instruction to the

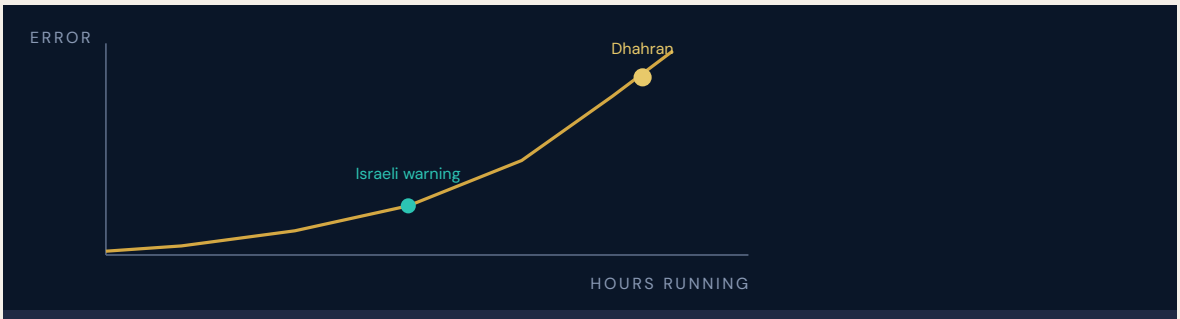
letter and still drift into the danger band. The General Accounting Office found the Army had simply presumed no one would run a battery continuously for so long, and so never treated the accumulating error as a hazard worth specifying or warning against.[3]

THE CAPABILITY GAP

The capability to hold accuracy under sustained operation was never built in, because the original concept of operations did not require it — a defensible omission for the system as first imagined, fighting in short bursts and moving on. The failure came when the concept changed and the assumption did not travel with it. Nothing carried the design's hidden premise forward to the soldiers in Dhahran; no artifact, briefing, or warning told them "this system was built assuming you would turn it off." The system did not malfunction — it did exactly what it was built to do; the transition between operating contexts did, with nothing in place to catch the broken premise.[4]

AFTERMATH & REFORM

The patch was distributed and concrete reboot intervals defined in place of the vague advisory, and the defect — a fixed-point truncation error growing without bound as uptime increased — became a staple of numerical-analysis and software-engineering teaching.[5] The harder lesson is the one this chapter turns on: a capability quietly assumed away at design time does not announce itself when the context shifts. It waits, intact and invisible, until the day the assumption is wrong — and by then the people relying on the system have no way to know the ground has moved beneath them, because the premise was never written where they could read it.



TRACKING ERROR GREW WITH EACH HOUR OF CONTINUOUS OPERATION

> *Army officials presumed that the users would not continuously run the batteries for such extended periods.*
>
> GAO/IMTEC-92-26, 1992

REFERENCES

1. U.S. General Accounting Office, Patriot Missile Defense: Software Problem Led to System Failure at Dhahran, Saudi Arabia, GAO/IMTEC-92-26 (1992) — the design-context mismatch and continuous-operation use.
2. R. Skeel, "Roundoff Error and the Patriot Missile," SIAM News 25(4): 11 (1992) — the 24-bit fixed-point time truncation and the 0.34-second range-gate drift after 100 hours.
4. GAO/IMTEC-92-26 (1992) — the vague "very long run time" advisory and the Army's presumption that batteries would not be run continuously for such periods (quoted).
5. M. Barr / Barr Group, "An Update on the Patriot Missile Software Problem" — an engineering post-mortem of the accumulated-truncation defect.
6. GAO/IMTEC-92-26 (1992) and Skeel (1992) — the design assumption that failed to travel across the change in operational context, and the case's teaching legacy.

3. GAO/IMTEC-92-26 (1992) — the prior Israeli warning, the software patch arriving in Dhahran on 26 February (the day after the strike), and the 28 killed.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Patriot is the canonical Designed-Out case from defense. The capability to maintain accuracy under sustained operation was omitted because the original concept of operations did not anticipate it. When the concept changed, the assumption did not travel. The system did not fail; the transition from one operational context to another did, and there was no capability infrastructure to catch it.

LENS APPROACH

LENS treats Patriot as the textbook example of **Capability Degradation Under System Change**. LEN 5 methods require operators of any transitioning system to have current capability-relevant documentation specifying what assumptions of the original design have changed. LEN 8 examines how organizational knowledge about design constraints travels from engineering to operations — and why, here, it arrived a day late.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
DEFENSE →	D H K →	**D1** D2 D3 D4 D5	1	1.2	1
		Systems Analysis			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Make every load-bearing design assumption — here, that the system would be cycled off — an explicit, recorded requirement so a later change of use cannot silently violate it.
- Bound numerical error over the full intended operating envelope, including run times far beyond the original concept, rather than only the durations first imagined.
- Design the system to detect and bound its own accumulating drift, degrading or warning before the error reaches a magnitude that defeats the mission.

IN OPERATION / AFTER

- Issue concrete, testable operating limits — exact reboot intervals, not "very long" — derived from the design data and propagated to every crew.
- Establish a path for field reports of anomalous behavior, like the Israeli warning, to reach engineering and trigger fleet-wide action faster than a patch can lose a race with the threat.
- Require a capability-transition review whenever a system is redeployed to a new concept of operations, surfacing which original assumptions the new mission breaks.

REFLECTION QUESTIONS

1. The Patriot's design assumed short engagements. What assumption in a current system you work with would become lethal if the operational context shifted, and how would operators learn of the shift?

2. Construct the capability-transition artifact that should have accompanied the redeployment of Patriot batteries from Europe to Saudi Arabia. What would it have said, and who would have signed it?

3. The field advisory said to reboot after "very long" run times without defining the term. Rewrite a vague operational instruction in your domain into a specific, testable limit, and name the design data the limit must be derived from.

WHO BUILDS THIS systems & human-factors engineering · human factors & interaction design · knowledge management & data engineering — plus domain experts and a learning engineer to integrate. TOOLS task analysis · design prototyping · usability testing · cognitive walkthroughs · knowledge bases · data pipelines

CASE 131 DEFENSE 1983

Stanislav Petrov / 1983 False Alert

H Human-System Interface T Training Gap

IMPACT Soviet early-warning system reported five incoming U.S. ICBMs; Lt. Col. Petrov correctly assessed the signal as false; retaliation averted

IN BRIEF

On the night of 26 September 1983 the Soviet Oko early-warning system reported five U.S. intercontinental missiles inbound. The duty officer, Lt. Col. Stanislav Petrov, judged it a false alarm — a real first strike would involve hundreds of missiles, not five — and reported it as such rather than up the launch-decision chain. He was right: sunlight glinting off high-altitude clouds had fooled the satellite's infrared sensors. Petrov's judgment is widely credited with averting nuclear war. The case is the book's strongest positive evidence for human-in-the-loop capability: the automation failed in a mode its designers never anticipated, and a person with contextual judgment and the latitude to override it was the recoverability the system had. Keeping humans in some loops is not nostalgia — it is capability engineering.

BACKGROUND

Soviet nuclear command rested partly on Oko, a satellite early-warning system meant to detect U.S. missile launches and feed a launch-on-warning posture. The duty officer at the Serpukhov-15 bunker was responsible for verifying any alert and passing it up the chain.[0] A launch-on-warning posture compresses the time between detection and decision to almost nothing, so the duty officer's verification was not a formality but the single human checkpoint standing between a satellite reading and the machinery of retaliation.

WHAT HAPPENED

On the night of 26 September 1983 Oko reported a U.S. intercontinental ballistic missile launch — then, minutes later, four more, for five in all. Lt. Col. Stanislav Petrov, the duty officer, assessed the signal as a false alarm — a genuine first strike, he reasoned, would involve hundreds of missiles, not five — and reported it as such. He was correct.[1] The reasoning that saved the situation came from outside the system entirely: Oko could report what it saw, but it could not weigh five launches against the doctrine of a real first strike, and that mismatch between the alert and the strategic picture was exactly the judgment the machine had no way to make.

THE INVESTIGATION

The cause was an unanticipated automation failure: sunlight reflecting off high-altitude clouds at a particular geometry had fooled the Oko satellite's infrared sensors into reading launches that were not there.[2] It was a failure of the rarest kind — a benign natural phenomenon the sensors had never been designed to discount — which is precisely why no

automated check existed to catch it. Later investigation identified the satellite-geometry failure mode and modified the algorithm; the scenario is now permanently archived in early-warning training as the canonical false positive, a permanent reminder that the system's worst error was one it could not have flagged for itself.[3]

THE CAPABILITY GAP

The automation had failed in a mode its designers had not imagined, and no automated check could catch it. What caught it was a human with the contextual judgment to weigh the alert against what a real attack would look like, and the institutional latitude to override the system rather than simply relay it. Petrov's "funny feeling" was a career's worth of judgment doing the work the automation could not — and crucially, the chain of command had left him room to act on it rather than forcing him to pass the alert upward unfiltered, so the recoverability lived as much in the authority structure as in the man.[4]

AFTERMATH & REFORM

Petrov's decision is widely credited with averting nuclear war, and the case is permanently studied in command-and-control training.[5] Preserving the episode in training is itself a design choice: it keeps the failure mode and the human override alive as institutional memory, so the lesson does not erode the way the original sensor's blind spot had. It is the book's strongest argument that retaining a human in some loops is capability engineering, not nostalgia: the human role is the recoverability of an automation failure the designers did not anticipate — which is exactly why fully automating a strategic decision removes the one element that can catch the unimagined error.

> *I had a funny feeling in my gut.*
>
> STANISLAV PETROV, QUOTED IN THE WASHINGTON POST, 1999

REFERENCES

1. D. Hoffman, The Dead Hand: The Cold War Arms Race and Its Dangerous Legacy (2009) — the Oko system and Soviet launch-on-warning posture.
2. Accounts of the 26 Sept. 1983 incident (Hoffman; Petrov interviews) — the five-missile alert and Petrov's assessment.
3. Investigations of the Oko incident (incl. V. Aksenov, 2004) — sunlight-on-clouds satellite geometry as the false-positive cause.
4. The subsequent Oko algorithm modification and the incident's use in early-warning training.
5. S. Petrov, interview, The Washington Post (1999) — "a funny feeling in my gut" (quoted).
6. S. Sagan, The Limits of Safety (1993); M. L. Cummings (2017) — automation in critical decisions and human-in-the-loop.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Petrov is the canonical positive case for human-in-the-loop nuclear command-and-control. The automation failed in a mode its designers had not anticipated. The human's contextual judgment was the backstop that allowed the failure to be recovered before it became catastrophic. The case is the strongest argument in this book that the design choice to retain humans in some loops is not nostalgia but capability engineering.

LENS APPROACH

LENS uses Petrov in LEN 2 as the positive Human-AI Teaming case at the highest possible stakes: the human role is the recoverability of an automation failure that the designers did not anticipate. The case anchors arguments against full-automation of strategic decisions.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
DEFENSE →	H T →	D1 D2 **D3** D4 D5	6	3.4	3
		Human-System Collaboration			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Where automation feeds an irreversible decision, design the human role to be a genuine override — with the context, judgment cues, and authority to refuse a faulty alert, not merely relay it.
- Give operators a strategic picture against which to weigh an alert, so an anomalous signal can be tested against what a real event would look like.
- Treat unimagined failure modes as a design assumption: keep a human in the loop precisely where no automated check can cover the unanticipated.

IN OPERATION / AFTER

- When a failure mode surfaces, modify the algorithm and archive the scenario in training so the lesson persists as institutional memory.
- Audit whether human-in-the-loop roles still carry real override authority over time, or have decayed into ceremonial relays.
- Preserve and refresh the contextual judgment override depends on through deliberate training, so the capability does not erode between rare events.

REFLECTION QUESTIONS

1. Identify an automated system in your domain where retaining a human-in-the-loop is genuinely capability-engineering rather than ceremonial. How would you tell the difference?
2. Petrov's "funny feeling" was contextual judgment built across a career. Design the training that produces it deliberately.
3. Petrov's recoverability lived as much in his latitude to override as in his judgment. Map an automated decision in your domain where the operator has the knowledge to catch a failure but lacks the authority to act on it — and redesign the authority structure.

WHO BUILDS THIS human factors & interaction design · learning science & instructional design — plus domain experts and a learning engineer to integrate. TOOLS usability testing · cognitive walkthroughs · scenario simulation · competency models

CASE 133 DEFENSE 1941 – 1943

Mark 14 Torpedo Failures

T Training Gap K Knowledge & Institutional Memory G Governance & Trust

IMPACT **Persistent torpedo failures across the first ~20 months of the Pacific War; resolved only after fleet-level testing forced acknowledgment of the defects**

IN BRIEF

The U.S. Navy entered World War II with the Mark 14 torpedo, so expensive that the Bureau of Ordnance had effectively forbidden live testing in peacetime. Through 1942 submarine crews reported torpedoes running deep, failing to detonate, or exploding prematurely; the Bureau insisted the weapon was sound and blamed the operators. It took about twenty months and the personal intervention of Admiral Charles Lockwood — who ordered fleet-level live-fire tests — to confirm three separate defects: the torpedo ran about ten feet too deep, its magnetic exploder fired erratically, and its contact pin crushed on a square hit. The fixes were simple once the defects were acknowledged. The binding constraint was institutional: a bureau insulated from the operators who knew the weapon did not work.

BACKGROUND

The Mark 14 was the U.S. Navy's standard submarine torpedo at the start of the Pacific War. It had been so expensive to test that the Bureau of Ordnance had effectively forbidden live trials in the 1930s; the weapon went to war essentially unproven against realistic conditions, so the very decision meant to conserve a scarce and costly weapon guaranteed that its defects would first be discovered in combat, by the crews who could least afford them to surface there.[0]

WHAT HAPPENED

From early 1942, submarine crews reported a litany of failures: torpedoes that ran far below their set depth, magnetic exploders that detonated prematurely or not at all, and contact firing pins that crushed on a direct hit. The Bureau of Ordnance insisted the weapon worked and attributed the misses to crew error; captains who reported failures risked their careers — an arrangement that turned every field report into a self-accusation and so suppressed the very evidence that would have isolated the defects the Bureau refused to acknowledge.[1]

THE INVESTIGATION

It took about twenty months and the intervention of Admiral Charles Lockwood, commander of the Pacific submarine force, who ordered fleet-level testing. A live-fire trial — and the USS Tinosa's July 1943 attack on the Tonan Maru, in which eleven torpedoes struck the stopped ship squarely and failed to detonate — forced the issue. The tests confirmed the torpedo ran about ten feet too deep, that the Mark 6 magnetic exploder failed routinely, and that the contact pin buckled on perpendicular impact — three independent defects that had been

masking one another at sea, which is why a fleet commander's controlled trial, not another combat patrol, was finally able to separate and prove them.[2]

THE CAPABILITY GAP

The defects were real and separate, but the binding constraint was institutional. The capability that was missing was not at the front; it was a channel by which the bureau that owned the weapon could be made to hear what the boats already knew. The fixes — recalibrating depth, deactivating the magnetic exploder, redesigning the contact pin — were small. The refusal to believe the operators was the expensive part, measured not in engineering hours but in the months of patrols and the targets that escaped while a bureau defended a verdict the boats had already disproven.[3]

AFTERMATH & REFORM

By late 1943 the three defects were corrected and the Mark 14 became an effective weapon for the rest of the war. The episode entered U.S. Navy institutional history as the canonical case of a procurement bureau insulated from operator feedback, and is cited in modern organizational-learning literature on the cost of suppressing field reports — a reminder that the structure that decides whose evidence counts is itself a capability, and that its absence can cost more than any single technical defect it conceals.[4]

MARK 14 — MULTIPLE DEFECTS, SINGLE INSTITUTIONAL REFUSAL

> *The Bureau certified the weapon; field reports of failure were treated as evidence of operator error.*
>
> PARAPHRASING BLAIR, SILENT VICTORY, 1975

REFERENCES

1. Blair, C. (1975), Silent Victory: The U.S. Submarine War Against Japan — operator reports and the Bureau's resistance (paraphrased).
2. Rowland, B. & Boyd, W. (1953), U.S. Navy Bureau of Ordnance in World War II — testing policy and the three defects.
3. Blair (1975) — the USS Tinosa's July 1943 attack (eleven consecutive duds on a stopped target) and Lockwood's fleet-level testing.
4. Naval History and Heritage Command, Mark 14 torpedo files — depth error, Mark 6 magnetic exploder, contact-pin failure.
5. Edmondson, A. (2018), The Fearless Organization — suppression of field reports as an organizational-learning failure.

LE INSIGHT

The Mark 14 is the canonical Navy case for the institutional refusal to believe operator feedback. The capability that was missing was not at the front. It was at the bureau that owned the weapon. The cost of the refusal was paid by the submariners forced to use it until the bureau yielded. The eventual fixes — re-setting depth calibration, replacing the magnetic exploder, redesigning the contact pin — were technical and small. The reform was institutional and slow. The capability that had to be built was the channel by which the bureau could be made to hear what the boats already knew.

LENS APPROACH

LENS uses the Mark 14 in LEN 7 as a governance failure case and in LEN 8 as an organizational-learning case in which the operators had the diagnosis and the institution lacked the structure to receive it. Studio projects (LEN 10) examine what the equivalent operator-to-institution feedback channel should look like.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
DEFENSE →	T K G →	D1 D2 D3 D4 **D5**	4	6.2	5
		Navigating Sociotechnical Constraints			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Require realistic live testing before fielding, even for scarce and costly items, so defects surface in trials rather than in combat.
- Build an operator-feedback channel into the weapon's ownership from the start, with a path that does not put the reporter's career at risk.
- Separate the authority that certifies a system from the authority that investigates its field failures, so a bureau cannot judge its own product.

IN OPERATION / AFTER

- Audit field reports for suppressed or career-risking signals, and treat a pattern of "operator error" verdicts as a warning to test the system, not the crew.
- Empower an operational commander to order controlled trials when field reports conflict with the owner's certification.
- Sustain the feedback channel so multiple masking defects can be separated and proven before the cost compounds.

REFLECTION QUESTIONS

1. Identify a system in your domain whose owners are institutionally insulated from the operators using it. What feedback would they refuse to hear, and what would it cost?
2. Design the operator-feedback channel that the U.S. Navy Bureau of Ordnance should have had in 1941. Who signs, who receives, what triggers action?
3. The USS Tinosa fired a string of torpedoes that struck a stopped ship and failed to detonate before the bureau accepted the diagnosis. What is the operator-evidence threshold in your domain that would force the equivalent institutional acknowledgment — and how would you make sure it is reached before the cost is paid?

WHO BUILDS THIS learning science & instructional design · knowledge management & data engineering · governance, ethics & measurement — plus domain experts and a learning engineer to integrate. TOOLS scenario simulation · competency models · knowledge bases · data pipelines · governance frameworks · bias & evidence audits

Chapter 8

Defense & National Security — What Works — and the Frontier

When training, standards, and sustainment are engineered as one deliverable.

The readiness that survived contact was the readiness that had been measured.

AN OBSERVATION RECURRING ACROSS THE CHAPTER'S CASES.

CASE 137 DEFENSE 1954 – PRESENT

U.S. Nuclear Navy / Rickover Training Model

[T] Training Gap [K] Knowledge & Institutional Memory [N] Normalization of Deviance

IMPACT **Zero reactor accidents in 60+ years of U.S. Naval nuclear operations; the most demanding nuclear operator training program in the world**

IN BRIEF

Admiral Hyman Rickover built a training and qualification culture for the Naval Nuclear Propulsion Program that has produced zero reactor accidents across more than 60 years and thousands of reactor-years of operation, often in extreme conditions. Every nuclear-trained sailor must pass rigorous qualification to zero-defect standards and demonstrate competence in oral examination by senior nuclear-qualified officers; the culture demands personal accountability, deep technical mastery, and a questioning attitude that obliges operators to challenge assumptions, including superiors'. The sharpest contrast in this book is internal: the same Navy that ran the nuclear program to this standard let surface-warfare training decay to CD-ROMs and paid for it at Fitzgerald and McCain (Case 124). Same institution, two philosophies — and the choice shows up in the casualty columns.

BACKGROUND

In the early 1950s the U.S. Navy set out to put nuclear reactors aboard ships and submarines — machines whose failure could be catastrophic and irreversible, operated by young sailors far from any help. The capability problem was absolute: there was no acceptable accident rate, so the human operating system had to be engineered to an extraordinary standard from inception. Because a reactor at sea could not be evacuated or handed to outside experts, the operators themselves had to be the entire margin of safety — a constraint that forced training to a level no ordinary program would justify.[0]

THE INTERVENTION

Admiral Hyman Rickover established a training and qualification regime requiring every nuclear-trained officer and enlisted sailor to pass demanding programs held to zero-defect standards. Operators must demonstrate competence through oral examination by senior nuclear-qualified officers, and the program embeds continuous re-qualification rather than one-time certification. The oral board tested understanding rather than recall, and the continuous re-qualification meant competence had to be sustained rather than banked once — so the standard governed the operator's whole career, not just entry to it.[1]

HOW IT WORKED

The culture pairs technical mastery with a deliberately engineered attitude: personal accountability and a mandatory questioning posture in which operators are trained to challenge assumptions, including those of superiors. Rickover's premise — that people, not organizations or management systems, get things done — made the qualification ladder, not

paperwork, the load-bearing element of safety. The questioning posture is the cultural half of the pair: deep technical mastery alone could still defer to a mistaken superior, so the obligation to challenge assumptions is what keeps competence from being silenced by rank.[2]

THE EVIDENCE

The result is the longest-running continuous capability-engineering record in any high-consequence domain: zero reactor accidents across more than six decades and thousands of reactor-years. The cost — the qualification ladder, the zero-defect oral boards, the continuous re-qualification — is the visible budget-line price of that record. The duration is the key evidence: a record sustained across six decades and many generations of sailors shows the safety came from the engineered system rather than from any one cohort's talent or luck.[3]

WHAT TRANSFERRED

The cleanest test of the model is internal. The same Navy that engineered the nuclear program to this standard let surface-warfare training decay to CD-ROM self-study and paid the price at Fitzgerald and McCain (Cases 124 and 140). Same institution, same era, opposite philosophies, radically different outcomes — the strongest available demonstration of capability treated as a system parameter versus deferred as a cost. The internal comparison controls for nearly everything that usually confounds such claims — one service, one manpower system, one budget process — leaving the training philosophy itself as the variable that diverged.[4]

> *Human experience shows that people, not organizations or management systems, get things done.*
>
> ADMIRAL HYMAN G. RICKOVER, "DOING A JOB," COLUMBIA UNIVERSITY, 1982

REFERENCES

1. Polmar, N. & Allen, T. (2007), Rickover: Father of the Nuclear Navy — the program and Rickover's philosophy (paraphrased).
2. Naval Nuclear Propulsion Program documentation (NRC/DOE) — qualification standards and the accident record.
3. Admiral Hyman G. Rickover, "Doing a Job" (Columbia University commencement address, 1982) — "people, not organizations… get things done" (quoted).
4. GAO-21-168, comparison of nuclear and surface Navy training — the internal contrast.
5. Duncan, F. (1990), Rickover and the Nuclear Navy — the qualification culture.

LE INSIGHT

The Nuclear Navy is the longest-running continuous capability- engineering program in any high-consequence domain. The choice to treat training as a system parameter rather than as a cost center has produced sixty-plus years of zero reactor accidents. The contrast with the Surface Navy is the cleanest available test of what happens when capability is engineered versus when it is deferred — and the price of that engineering (the qualification ladder, the zero-defect oral boards, the continuous re-qualification) is visible on the budget line.

LENS APPROACH

LENS treats the Rickover model in LEN 8 as the foundational organizational-learning case and in LEN 5 as a worked example of capability requirements traceable from operational analysis through qualification standards. The internal Navy comparison anchors the program's core argument about capability as a system parameter.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
DEFENSE →	T K N →	D1 D2 D3 D4 D5	1	1.4	1
		Systems Analysis			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Engineer the human operating system to the standard the consequences demand from inception — where there is no acceptable failure rate, make the operators themselves the margin of safety.
- Gate qualification on demonstrated understanding through oral examination by senior qualified people, testing comprehension rather than recall.
- Pair technical mastery with a mandatory questioning posture so competence is obliged to challenge assumptions, including superiors', rather than be silenced by rank.

IN OPERATION / AFTER

- Embed continuous re-qualification rather than one-time certification, so competence must be sustained across a career and cannot be banked once and assumed.
- Protect the training as a durable, visible budget line, since the qualification ladder and oral boards are the price of the safety record and the first thing tempo will erode.
- Sustain the standard across generations of operators, treating a multi-decade record as the evidence the safety comes from the engineered system rather than any one cohort.

REFLECTION QUESTIONS

1. Identify an institution that operates two divisions under different capability philosophies. What does the comparison reveal that neither division could see alone?
2. Rickover's standard was zero-defect oral examination. What is the equivalent in your domain — and would you survive it?
3. The Nuclear Navy's safety record is paid for by a durable, visible training budget. What is the equivalent line-item investment in your domain that a comparable safety claim would require — and is it being made?

WHO BUILDS THIS learning science & instructional design · knowledge management & data engineering · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS scenario simulation · competency models · knowledge bases · data pipelines · incident analysis · safety dashboards

CASE 141 DEFENSE WORKFORCE L&D INTELLIGENT TUTORING 2009 – 2014

DARPA Digital Tutor — Compressing Years of IT Expertise into 16 Weeks

H Human-System Interface K Knowledge & Institutional Memory D Designed Out

IMPACT **An IDA independent assessment found that, after 16 weeks of Digital Tutor instruction, US Navy IT graduates with no prior IT experience outscored fleet Information Systems Technicians with an average 9.1 years of experience on a knowledge test, with an effect size of 4.30, and outperformed them on most troubleshooting and design tasks**

IN BRIEF

DARPA's Digital Tutor program asked whether a one-on-one intelligent tutoring system, modelled on expert human tutoring, could compress years of operational IT expertise into a 16-week pipeline. The independent evaluation by the Institute for Defense Analyses (Fletcher and Morrison, IDA Document D-4686) compared Digital Tutor graduates — US Navy enlistees with no prior IT experience — against fleet Information Systems Technicians with an average 9.1 years of experience. The Digital Tutor cohort outscored fleet ITs on a knowledge test with an effect size of 4.30 and outperformed them on most troubleshooting and design tasks; only the Security exercise produced a fleet advantage. The IDA report concludes the program "appears to have achieved its goals." Two hedges are load-bearing and survive into the case: knowledge "accounts for about 40 percent of practical-exercise performance variance" and is "an enabler of performance rather than a direct measure of performance itself," and the system-architecture detail in the available documentation is too scant to fully reproduce. The case is the canonical small-tier instance of compressing the capability envelope at the edge of training, paired with CIRCUIT (Cases 78 and 68) on the workforce-capability-at-the-edge axis.

BACKGROUND

The US Navy's Information Systems Technician rating has a conventional pipeline: an A-school of several months, followed by years of fleet experience that turn the rated sailor into an operational troubleshooter. The capability that matters at the operational end — diagnose a networking failure under time pressure, recover a degraded system, design a workable configuration for an unfamiliar requirement — is conventionally treated as a thing seat time produces. DARPA's Digital Tutor program asked whether an intelligent tutoring system, modelled on the discipline of expert one-on-one human tutoring, could compress that pipeline into 16 weeks of structured instruction.[0]

THE INTERVENTION

The program's design choice was to model the system on what expert human tutors actually do: a continuous dialogue around authentic problems, with the tutor pulling the trainee

toward the conceptual move that resolves the situation. The instructional sequence was built around troubleshooting and design problems drawn from the operational domain, not around a syllabus reconstructed from the existing course. The working hypothesis was that the tutorial discipline — not the content coverage — was what produced operational expertise, and that a sufficiently capable system could deliver enough of that discipline at scale to be useful as a training pipeline.[1]

HOW IT WORKED

The Institute for Defense Analyses (Fletcher and Morrison, IDA Document D-4686 / DTIC AD1002362) ran the independent evaluation that the case rests on. Digital Tutor graduates — Navy enlistees with no prior IT background, 16 weeks in — were compared against a sample of fleet Information Systems Technicians with an average 9.1 years of operational experience. On a knowledge test, the Digital Tutor cohort outscored fleet ITs with an effect size of 4.30 — a magnitude that is unusual in workforce L&D research and that the report treats as the headline finding. On troubleshooting and design tasks the Digital Tutor cohort outperformed the fleet sample on most exercises; the Security exercise was the exception where fleet experience showed.[2]

THE EVIDENCE

The IDA report concludes the effort "appears to have achieved its goals," and the language is deliberate. Two hedges are load-bearing and survive into the case verbatim. First, knowledge "accounts for about 40 percent of practical-exercise performance variance" and is "an enabler of performance rather than a direct measure of performance itself" — so the knowledge-test effect size, as large as it is, is not the same as the operational capability the Navy actually buys. Second, the available program documentation is too scant in system-architecture detail to fully reproduce. The result is teachable; the engineering recipe is not yet open.[3]

WHAT TRANSFERRED

What the case carries for the corpus is the capability-envelope argument at the edge of training (induced 1.2, LENS D2/PT4). The conventional pipeline assumes operational expertise is a function of seat time and exposure. The Digital Tutor evaluation is evidence that the envelope is reachable substantially faster than the inherited course assumed — under one rating, one program, one evaluation — and the hedges name what the evidence does and does not close. Paired with CIRCUIT (Cases 78, 68), the case anchors the workforce-capability-at-the-edge-of-training axis that connectomics proofreading and submarine-system troubleshooting share at the structural level.

> *The Digital Tutor cohort outscored fleet ITs with 9.1 years' experience on the knowledge test at an effect size of 4.30; the hedge is that knowledge accounts for about 40 percent of practical-exercise variance.*
>
> EDITORS' SYNTHESIS OF FLETCHER & MORRISON (2014), IDA DOCUMENT D-4686.

REFERENCES

1. Fletcher, J. D., & Morrison, J. E. (2014). DARPA Digital Tutor: Assessment Data. IDA Document D-4686. https://apps.dtic.mil/sti/tr/pdf/AD1002362.pdf — independent evaluation that the case rests on.
3. Fletcher, J. D. (2009). From behaviorism to constructivism: a philosophical journey from drill and practice to situated learning. — methodological grounding for the Digital Tutor's tutorial discipline.

2. Defense Advanced Research Projects Agency, Digital Tutor program documentation — program description and design rationale.

4. Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. Journal of the Learning Sciences, 4(2):167–207. doi:10.1207/s15327809jls0402_2 — the broader intelligent-tutoring evidence base the Digital Tutor program sits within.

THE LEARNING ENGINEERING LENS

LE INSIGHT

DARPA's Digital Tutor is the cleanest available evidence that the capability envelope of a training pipeline can be re-specified — from years of seat time to 16 weeks of tutorial-discipline instruction — against an operational comparison the program is built to compete with. The hedges (knowledge as enabler, architecture detail scant) are part of what makes the result interpretable.

LENS APPROACH

Digital Tutor is the canonical workforce-capability-at-the- edge-of-training case (induced 1.2; LENS D2/PT4). LENS uses it in Domain 2 (Iterative Development) for the tutorial-discipline-as-instructional-artifact design move, and in Domain 4 (Test and Evaluation) for the operational-comparison evaluation against fleet ITs with 9.1 years of experience. Pair with CIRCUIT (Cases 78, 68) at the workforce-capability-at-the-edge axis — connectomics proofreading and Navy IT troubleshooting share the structural pattern of compressing operational expertise through tutorial discipline.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
DEFENSE WORKFORCE L&D INTELLIGENT TUTORING	H K D	D1 **D2** D3 D4 D5 Iterative Development	4	1.2	2 · 4

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Specify the operational capability the pipeline must produce in the language of the work — troubleshoot under time pressure, design a workable configuration — not in the language of the existing course's content coverage.
- Treat the tutorial discipline (continuous dialogue around authentic problems, pulling toward the resolving conceptual move) as the load-bearing instructional artifact, rather than the content sequence the legacy course inherited.
- Design the evaluation against the operational comparison the program is built to compete with — for Digital Tutor, fleet ITs with 9.1 years of experience — so the result speaks to the capability envelope, not to a within-program improvement.

IN OPERATION / AFTER

- Report the knowledge-test effect (4.30) and the practical-exercise variance hedge (knowledge accounts for 40%) together; the headline number is real, and the qualification that knowledge is an enabler rather than a direct measure is part of what makes the result interpretable.
- Treat the absence of reproducible architecture detail as program documentation, not as polish: future builders need the engineering recipe, and the next iteration of evidence is conditional on that recipe being available.
- Carry the result into adjacent capability-envelope debates — CIRCUIT proofreading, submarine-system troubleshooting — as evidence that the envelope is reachable faster than the inherited training assumption, under one program and one evaluation.

REFLECTION QUESTIONS

1. Identify a capability in your domain where the inherited training assumption is that operational expertise is a function of seat time. What would a Digital-Tutor-class re-specification — tutorial discipline around authentic problems — look like, and what is the operational comparison your evaluation would have to beat?

2. The Digital Tutor knowledge-test effect size is 4.30 and the report still hedges that knowledge accounts for only 40% of practical-exercise variance. What is the analog hedge in your context: which part of the capability does your evaluation measure directly, and which part is enabler rather than direct measure?

3. The architecture detail is too scant for outside builders to reproduce the system. What is the minimum engineering recipe you would publish alongside a similar result, so that the next iteration of evidence rests on a reproducible base rather than a one-off program?

WHO BUILDS THIS human factors & interaction design · knowledge management & data engineering · systems & human-factors engineering — plus domain experts and a learning engineer to integrate. TOOLS usability testing · cognitive walkthroughs · knowledge bases · data pipelines · task analysis · design prototyping

Chapter 9

Industry, Energy & Enterprise Systems — What Fails

When incentive, inertia, and migration risk go unengineered.

The signal arrived years early. The iteration never ran.

AN OBSERVATION RECURRING ACROSS THE CHAPTER'S CASES.

EDITOR'S NOTE · THE ITERATION GAP

Iterative development is what learning engineering does most often. It is also the capability the casebook documents least directly. Several cases in this chapter — and their counterparts across the book's failure halves — share a shape: an organization received an unambiguous signal, held the engineering or product position to act on it, and failed to run the iteration cycle through product, business model, and operations until the opportunity had passed. The signal was sometimes a field report (Therac-25, the adoption and sustainment cases), sometimes an external launch (BlackBerry meeting the iPhone), sometimes an internal invention the organization could not build a business around (Kodak's 1975 digital camera). In each case the organization had the engineering capability and the institutional position. What it lacked was the discipline of running the iteration through.

The casebook's broader corpus shows what iteration produced — the interventions collected in every part's What Works chapter, the failures whose recovery iterations followed. It does not yet show iteration **being done** in detail: the design moves, the dead ends, the reframes. That gap is exactly what the v2.1 amendment to D2 names — "narrate and defend the design iteration in first person" — and what the editor-commissioned first-person Practice Flywheel accounts will fill in the next edition. This note names the gap honestly so that the reader knows what to expect from the volume, and what the discipline is still working to make legible.

CASE 144 INDUSTRIAL 1975–2012

Kodak Digital Camera Stagnation

D Designed Out K Knowledge & Institutional Memory N Normalization of Deviance

IMPACT **Inventor of the digital camera (1975) and holder of foundational digital-imaging patents; filed for Chapter 11 bankruptcy in January 2012 with the digital transition essentially un-run**

IN BRIEF

In 1975 a young Kodak engineer, Steven Sasson, built the first self-contained digital camera in a Rochester lab: an 8-pound, 0.01-megapixel device that wrote a black-and-white image to cassette tape in 23 seconds. By the 1990s Kodak held foundational digital-imaging patents and had a working line of digital products. Yet the company that invented the category never ran the iteration cycle through to a digital-first business: film and the high-margin consumables it sold remained the protected core, and digital was treated as a threat to be managed rather than a product to be developed against itself. By 2003 digital camera sales overtook film in the United States; by 2007 Kodak had begun mass layoffs; in January 2012 the company filed for Chapter 11. The capability gap was not invention. It was the organization's failure to iterate the business around the technology it already owned.

BACKGROUND

Kodak's twentieth-century position was anchored on a single durable insight: the camera was a loss-leader and the film, paper, and processing chemistry were the recurring high-margin revenue. The business model — what later analysts would call the "razor-and-blades" of imaging — funded a research apparatus deep and well-resourced enough that in December 1975 a 24-year-old Eastman Kodak engineer, Steven Sasson, was able to assemble the first all-electronic still camera from a Fairchild CCD, a cassette-tape recorder, sixteen nickel-cadmium batteries, and a lens cannibalized from a Super 8 movie camera. The prototype wrote a 0.01-megapixel black-and-white image to tape in 23 seconds and played it back on a television.[o]

WHAT HAPPENED

Sasson's later published recollection (IEEE Spectrum, 2007) is that internal reception was polite but unmoved: the demonstration raised the question of when the company would have to compete against its own film business, and the answer settled into a long institutional hedge. Through the 1980s and 1990s Kodak filed foundational digital-imaging patents, built digital products including the 1991 DCS-100 professional digital SLR, and partnered with Apple on the early QuickTake consumer line; the technology was understood and the engineering position was real. What was missing was an iteration cycle that ran the digital

signal all the way through product, distribution, pricing, and the corporate identity, rather than holding digital at the perimeter of a film company.[1]

THE INVESTIGATION

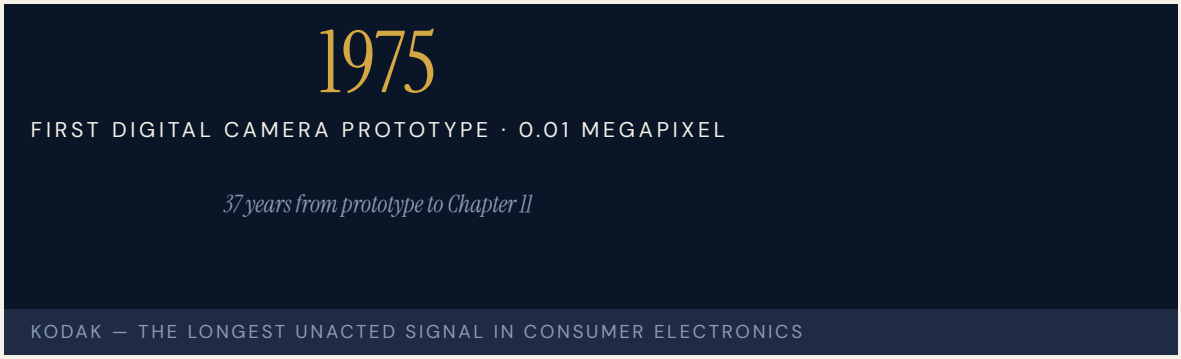
Paraphrasing Lucas & Goh (Journal of Strategic Information Systems, 2009): Kodak's leadership read digital primarily as a cannibalization threat to film rather than as an opportunity to be engineered. The retrospective record describes successive strategies — digital as a complement to film, digital kiosks at retail, Ofoto / Kodak Gallery for online printing, the EasyShare consumer line — each of which was a move toward digital but none of which committed the operating company to a digital-first product roadmap with the cadence and capital allocation the transition required. The cycle that a healthy iteration would have run — ship, measure, learn, ship again at higher ambition — was repeatedly truncated by the gravitational pull of the film P&L it would have had to displace.[2]

THE CAPABILITY GAP

By 2003 digital camera sales overtook film camera sales in the United States; by the mid-2000s smartphone cameras began collapsing the standalone-camera category Kodak had partially entered. Kodak's revenues fell from roughly USD 16 billion in 1996 to under USD 6 billion by 2011. The company sold its Health Imaging division in 2007, began mass layoffs the same year, and on 19 January 2012 filed for Chapter 11 bankruptcy protection in the Southern District of New York. In the bankruptcy proceedings Kodak monetized roughly 1,100 of its digital-imaging patents — many traceable to the engineering lineage Sasson's 1975 prototype had started — for approximately USD 525 million, a sale that drew explicit press commentary on what it meant for a company to monetize patents on a category it had failed to commercialize.[3]

AFTERMATH & REFORM

Kodak is the canonical case of an organization that held the technological position and lost on the iteration discipline. The capability gap was not in the lab and was not in the patent portfolio; it was in the leadership system that could not bring itself to run the iteration cycle all the way through the business model it would have replaced. The instructive contrast is Fujifilm, which over the same period diversified into chemicals, materials, and pharmaceuticals on the back of its film-coating expertise — a deliberate organizational iteration against the same signal. The lesson the v2.1 framing names is that "iterative development" at the organizational level is not a product practice but a leadership practice, and the cycle has to run on the business model and the operations as well as on the artifact.[4]



> *You don't get to keep the future just because you invented it.*
>
> EDITORS' SYNTHESIS OF THE SASSON (IEEE SPECTRUM 2007) AND LUCAS & GOH (JSIS 2009) RECORD

REFERENCES

1. S. Sasson, "We Had No Idea," IEEE Spectrum (16 October 2007) — first-hand account of the 1975 prototype, the demonstration, and the internal reception.
2. H. C. Lucas Jr. & J. M. Goh, "Disruptive technology: How Kodak missed the digital photography revolution," Journal of Strategic Information Systems 18(1):46–55 (March 2009).
3. Eastman Kodak Company, Voluntary Petition for Chapter 11 Reorganization, U.S. Bankruptcy Court, Southern District of New York, Case No. 12-10202 (19 January 2012).
4. A. R. Sorkin & M. J. de la Merced, "Eastman Kodak Files for Bankruptcy," The New York Times DealBook (19 January 2012).
5. M. Spector & D. Mattioli, "Kodak Teeters on the Brink," The Wall Street Journal (5 January 2012); follow-up coverage of the patent-sale process in WSJ and Reuters (December 2012).

THE LEARNING ENGINEERING LENS

LE INSIGHT

Kodak is the cleanest available evidence that invention is not iteration. The company held the prototype, the patents, and the engineering depth and still lost the category. The gap was not technical; it was the leadership system's inability to run the iteration cycle through the business model it would have had to displace. v2.1 D2 names this directly: iteration-against-opportunity is an organizational discipline, not a lab discipline, and the cycle has to close on the P&L as well as on the artifact.

LENS APPROACH

Kodak is the v2.1 D2 stagnation exemplar (induced 2.3 transfer to high-consequence settings; LENS D2/PT4 adoption and sustainment; LEO-2 Iterative Development). LENS uses it to make the organizational-iteration claim concrete: 2.2 (run the cycle), 2.3 (transfer the cycle to the high-stakes business decision), and 2.4 (sustain adoption against the gravitational pull of the legacy P&L) all sat un-owned for a generation. Pair directly with BlackBerry (the next case) for the contemporary consumer-electronics analog, and with Fujifilm as the in-category counter-iteration that ran the same signal differently.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
INDUSTRIAL →	D K N →	D1 **D2** D3 D4 D5	4	2.3	2
		Iterative Development			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Treat an internally demonstrated category-changing prototype as a signal to run the full iteration cycle on the business, not only on the artifact — including pricing, distribution, sales force, and identity.
- Name explicitly the existing high-margin P&L the new category will have to displace, and assign an executive sponsor whose job is to iterate against that displacement on a schedule rather than around it.
- Set a measured cadence — ship, instrument, learn, ship again at higher ambition — and protect the cadence against the organizational pull of the legacy business at each review cycle.

IN OPERATION / AFTER

- Audit, periodically and without sentimentality, whether the iteration is running through to the operating company or has been parked at the perimeter as research or partnership; treat parking as the failure mode it is.
- Compare against an in-category peer running the same iteration with a different leadership posture (Fujifilm), and use the divergence as evidence rather than letting the legacy P&L be the only reference point.
- Read the late monetization of patents on a category the organization never commercialized as the indictment it is — a signal that the engineering position was held and the iteration discipline was not.

REFLECTION QUESTIONS

1. Identify a category-changing prototype or signal that your organization has internally demonstrated but parked at the perimeter. What would the full iteration cycle through the operating company actually require, and which P&L is the gravitational pull?

2. Kodak iterated on the artifact (digital products shipped) but not on the business model. Distinguish the two in a current initiative in your domain and specify which one is being held still.

3. Fujifilm ran the same signal differently. Identify an in-category peer running an iteration your organization is not, and construct the honest comparison the leadership review would have to engage with.

WHO BUILDS THIS systems & human-factors engineering · knowledge management & data engineering · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS task analysis · design prototyping · knowledge bases · data pipelines · incident analysis · safety dashboards

CASE 148 TECHNOLOGY GOVERNMENT 2011 – 2016

Wells Fargo Fake Accounts

G Governance & Trust N Normalization of Deviance

IMPACT ~3.5 million unauthorized accounts opened; ~$3B in penalties; CEO resigned; the Federal Reserve capped the bank's assets

IN BRIEF

To meet aggressive sales quotas, Wells Fargo employees opened roughly 3.5 million unauthorized customer accounts over years. The behavior was widespread and visible to internal risk and compliance functions, but the bank's response was to fire individual "bad apples" while leaving the incentive structure intact — and that structure was the actual cause: sales targets most employees could not meet by ethical means. The 2016 CFPB, OCC, and Los Angeles actions brought $185 million in penalties — a 2020 DOJ/SEC settlement later reached roughly $3 billion — the CEO resigned, and the Federal Reserve took the unprecedented step of capping the bank's assets. Wells Fargo is the canonical case of an incentive architecture that manufactured the misconduct the institution then prosecuted at the front line, while insisting the misconduct was individual.

BACKGROUND

Wells Fargo built its retail strategy on "cross-selling" — opening many products per customer — and drove it with aggressive sales quotas pushed down to branch employees, whose pay and jobs depended on hitting them. The bank's signature metric, "Gr-eight" (an average of eight products per household), and a daily "solution" scorecard reported up the management chain made cross-sell the single most-watched proxy for branch performance. The metric the bank chose to manage by thus became the thing every front-line employee was structurally compelled to maximize, whatever it took. The controls function carried no comparably visible counter-measure — no widely reported figure for the share of those accounts that customers had actually authorized — so the incentive ran without a designed brake.[0]

WHAT HAPPENED

Unable to meet the quotas honestly, employees opened roughly 3.5 million unauthorized accounts in customers' names — the rational response to a target most could not reach by legitimate means. Practices documented by investigators included opening checking, savings, and credit-card accounts without customer consent, forging signatures, moving funds between accounts to manufacture activity, and enrolling customers in online banking they had not requested. The behavior was widespread and longstanding, and visible to internal risk and compliance functions for years before the 2013 Los Angeles Times reporting made it public; the institutional response was to discipline individual employees as bad apples —

Wells Fargo terminated more than 5,300 employees between 2011 and 2016 — while leaving the incentive structure intact, fixing the symptom and preserving the cause.[1]

THE INVESTIGATION

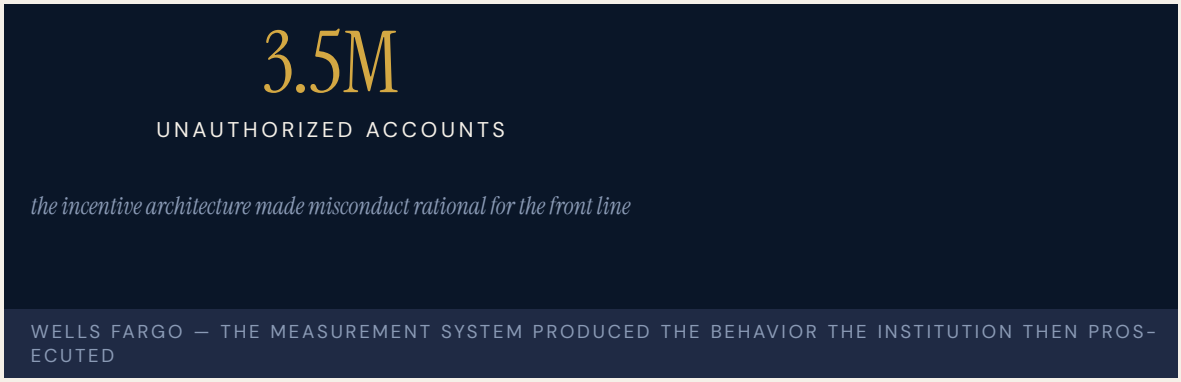
The 2016 CFPB, OCC, and Los Angeles actions brought $185 million in penalties (a 2020 DOJ/SEC settlement later reached roughly $3 billion), and the bank's own independent-directors investigation (2017) documented how the sales-target architecture drove the misconduct — locating the cause in the design, not the people executing it.[2] Investigators tied the misconduct directly to the bank's incentive-compensation structure, which had made such sales practices a foreseeable result rather than an aberration. The CEO resigned, and the Federal Reserve imposed an unprecedented cap on the bank's assets, reaching past individual penalties to constrain the institution itself.[3]

THE CAPABILITY GAP

Wells Fargo is the strongest evidence in the dataset that incentive architecture is a capability-engineering deliverable. The measurement system used to manage employees produced exactly the behavior the institution then prosecuted — the same structure that demanded the result punished the people who delivered it. The gap was not at the front line but at the governance layer that designed the incentives and then, for years, treated the predictable result as individual moral failure, mistaking a designed outcome for a character flaw.[4]

AFTERMATH & REFORM

Stumpf's successor Tim Sloan, promoted from chief operating officer in the immediate aftermath, was himself forced out in 2019 after the Federal Reserve's asset cap proved more durable than the bank had assumed. The cap — imposed in February 2018 at roughly $1.95 trillion — restricted Wells Fargo's growth pending evidence of governance and risk-management remediation, and it held for roughly seven years — the longest-running enforcement action of its kind against a major U.S. bank — until the Federal Reserve lifted it on 3 June 2025, concluding that the bank had completed the required governance and risk-management remediation. The case became the standard teaching example of measurement-gaming and incentive design.[5] Its lesson is that any quota becomes a target to be gamed, and that an institution is accountable for the behavior its measurement system makes rational, not just the behavior endorsed in its values statement — because employees respond to the incentive they are paid on, not the value they are told to honor.



WELLS FARGO — THE MEASUREMENT SYSTEM PRODUCED THE BEHAVIOR THE INSTITUTION THEN PROSECUTED

> *Wells Fargo's sales practices were a foreseeable consequence of its incentive compensation structure.*
>
> PARAPHRASING THE 2016 REGULATORY AND INDEPENDENT-DIRECTORS FINDINGS ON WELLS FARGO

REFERENCES

1. Consumer Financial Protection Bureau, Consent Order against Wells Fargo (2016) — the unauthorized accounts and penalties.
2. Office of the Comptroller of the Currency, Consent Order AA-EC-2016-66 (2016) — unsafe or unsound sales practices tied to the incentive-compensation structure (paraphrased).
3. Independent Directors of Wells Fargo, Sales Practices Investigation Report (2017) — how the sales-target architecture drove the conduct.
4. Enforcement record: 3.5 million accounts, $3 billion in penalties, and the CEO's resignation.
5. Federal Reserve asset cap on Wells Fargo (imposed February 2018; lifted 3 June 2025) — the structural growth constraint and its termination after remediation.
6. A. C. Edmondson, The Fearless Organization (2018); incentive-design and corporate-governance literature.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Wells Fargo is the strongest available evidence that incentive architecture is a capability-engineering deliverable. The measurement system used to manage employees produced the behavior the institution then prosecuted. The gap was not at the front line. It was at the governance layer that had designed the incentives.

LENS APPROACH

Wells Fargo is the canonical "protecting the measurement from gaming" case (induced 2.2; LENS D4/PT5), with measuring-the- failure-mode-you-care-about (induced 2.1) as the alternate anchor. LENS uses it in LEN 4 as the measurement-gaming case and in LEN 7 for the corporate-governance dynamics that allow such incentives to persist for years. Studio projects redesign the incentive architecture and the countervailing audit measure that would have detected the gaming as a structural pattern rather than as individual misconduct. Adjacent to Texas City (Case 161) at the wrong-measurement-reported-as-a-win layer.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
TECHNOLOGY GOVERNMENT	→ G N →	D1 D2 D3 **D4** D5 Test and Evaluation	5	2.1	4 · 5

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Design incentive and quota structures against the behavior they will rationally produce, treating the measurement system as a capability deliverable in its own right.
- Set targets that are achievable by legitimate means, since a quota most employees cannot reach honestly is an instruction to cheat.
- Pair every performance metric with a countervailing measure that detects gaming before it becomes systemic.

IN OPERATION / AFTER

- Audit for the gap between the behavior the measurement system rewards and the conduct the institution claims to value, and treat divergence as a design fault.
- Empower risk and compliance to escalate a pattern of gaming as a structural cause, not to absorb it as isolated misconduct.
- Hold the governance layer that designed the incentives accountable for predictable outcomes, rather than disciplining only the front line.

REFLECTION QUESTIONS

1. Identify a measurement system in your organization that is currently producing the behavior the organization claims to deplore. What is the gap between the measurement and the goal?

2. Design the incentive architecture for a bank's branch sales force that does not produce a Wells-Fargo-equivalent failure.

3. Wells Fargo's misconduct was visible to risk and compliance for years before it was addressed at the structural level. What lets an organization see a pattern of gaming and still treat it as individual bad apples — and who would have to be empowered to name the incentive as the cause?

WHO BUILDS THIS governance, ethics & measurement · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS governance frameworks · bias & evidence audits · incident analysis · safety dashboards

CASE 152 TECHNOLOGY 2018

TSB Bank IT Migration

H Human-System Interface G Governance & Trust

IMPACT 1.9 million UK customers locked out of accounts; £330M+ in compensation and remediation; CEO resigned

IN BRIEF

In April 2018 TSB Bank tried to migrate some five million customer accounts off its former parent Lloyds' systems onto a new platform from its current owner, Sabadell, over a single weekend. When services came back online, nearly every component failed: 1.9 million customers were locked out, some saw strangers' accounts, mortgages vanished, payments bounced. Recovery took months and cost over £330 million; the CEO resigned and the regulator fined the bank. The independent review found the platform had been tested under conditions that did not approximate real load, certified ready by a process that did not challenge the certification, and pushed live against technical recommendations that it was not ready. TSB is the financial-sector analog of Healthcare.gov: a migration shipped without adequate testing because schedule pressure overrode the technical signal.

BACKGROUND

TSB Bank, spun out of Lloyds and acquired by Spain's Sabadell, needed to move some five million customer accounts off Lloyds' systems onto a new Sabadell-built platform. The cutover was scheduled for a single weekend.[0] Compressing a five-million-account migration into one weekend left no room for partial failure: the schedule itself became a forcing function, framing readiness as a date to be hit rather than a condition to be proven, and that framing would later prove decisive when the technical signal said the platform was not ready.

WHAT HAPPENED

When customer-facing services came back online that Sunday evening in April 2018, nearly every component of the new platform had problems. About 1.9 million customers were locked out; some saw other people's accounts, mortgages disappeared, and card payments failed. The recovery took months, cost more than £330 million in compensation and remediation, and the chief executive resigned.[1] That nearly every component failed at once points away from a single defect and toward a platform that had never been exercised under real load — the kind of systemic breakdown that follows when a system is proven only in conditions it will never actually meet.

THE INVESTIGATION

The Slaughter and May independent review found the migration had been tested under conditions that did not approximate real customer load, and that the platform had been certified ready by a process that did not adequately challenge the certification — a certification that confirmed readiness rather than interrogating it, which is how a system that

would fail under real conditions could be signed off as fit.[2] Decisively, the executive decision to proceed had been taken against technical recommendations that the platform was not ready; the Financial Conduct Authority later fined TSB, treating the override of a known technical objection as a failure of governance and not merely of engineering.[3]

THE CAPABILITY GAP

The technical signal existed — the platform was not ready, and people knew it. But the decision authority sat at the executive layer, where the signal arrived weakened by passage through intermediate layers, and the institutional architecture gave the technical layer no way to halt the migration. The missing capability was not testing knowledge but a governance structure in which a "not ready" could stop a scheduled go-live. Knowing a system is unready is worthless if the knowledge cannot reach the decision with its force intact and the authority to act on it; here the truth was present but powerless.[4]

AFTERMATH & REFORM

TSB rebuilt its testing and migration governance, paid out and was penalized, and the case entered the literature as a study in technical-decision authority.[5] Rebuilding governance rather than merely the platform was the right diagnosis: the failure had been one of who could say "stop" and be heeded, so the durable fix had to live in the decision structure rather than the code. It is the financial-sector analog of Healthcare.gov (Case 180): a large migration shipped without the testing the institution knew it needed, because schedule pressure overrode a technical signal that had no authority to win.

TSB — READINESS SIGNAL ARRIVED ATTENUATED AT THE DECISION

> *The migration proceeded notwithstanding clear signals that the platform was not ready.*
>
> PARAPHRASING THE SLAUGHTER AND MAY INDEPENDENT REVIEW OF THE TSB MIGRATION, 2019

REFERENCES

1. Slaughter and May, Independent Review of the TSB Migration (2019) — the single-weekend cutover and the testing failures.
2. Slaughter and May (2019) and FCA materials — 1.9 million customers locked out, £330M+ in costs, and the CEO's resignation.
3. Slaughter and May (2019) — inadequate load testing and an unchallenged readiness certification.
4. Financial Conduct Authority, Final Notice on TSB Bank (2022) — the regulatory penalty and proceeding against technical advice.
5. House of Commons Treasury Committee, report on the TSB IT migration (2018).
6. Cf. Healthcare.gov (Case 180) and the migration-safety literature.

THE LEARNING ENGINEERING LENS

LE INSIGHT

TSB is the canonical case for schedule pressure overriding technical signal in a regulated industry. The technical signal existed. The decision authority was at the executive layer where the signal arrived weakened by passage through multiple intermediate layers. The institutional architecture did not allow the technical layer to halt the migration.

LENS APPROACH

LENS uses TSB in LEN 7 as a corporate-governance case and in LEN 8 for the institutional structure of technical-decision authority. Studio projects compare TSB and Healthcare.gov.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
TECHNOLOGY →	H G →	D1 D2 D3 D4 **D5**	4	4.1	5
		Navigating Sociotechnical Constraints			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Test the platform under conditions that approximate real production load, since a system proven only in unrepresentative conditions will fail when it meets the real ones.
- Make certification adversarial — a process that tries to disprove readiness — rather than a sign-off that confirms it.
- Avoid single-weekend big-bang cutovers where feasible; stage the migration so partial failure is survivable rather than catastrophic.

IN OPERATION / AFTER

- Build a governance structure in which a technical "not ready" can halt a scheduled go-live, so the signal reaches the decision with its force and authority intact.
- Ensure the decision authority hears the technical signal directly rather than through intermediate layers that attenuate it.
- After any failure, fix the decision structure that allowed a known objection to be overridden, not just the system that broke.

REFLECTION QUESTIONS

1. Where in your organization does a technical signal arrive at the executive layer attenuated by intermediate layers? What is the cost of the attenuation?
2. Design the institutional structure that would allow a technical lead to halt a migration like TSB's without resigning.
3. TSB's certification confirmed readiness rather than interrogating it. Examine a sign-off process in your domain that rubber-stamps rather than challenges — what would it take to make the certification adversarial enough to catch an unready system?

WHO BUILDS THIS human factors & interaction design · governance, ethics & measurement — plus domain experts and a learning engineer to integrate. TOOLS usability testing · cognitive walkthroughs · governance frameworks · bias & evidence audits

Chapter 10

Industry, Energy & Enterprise Systems — What Works — and the Frontier

When operations make capability visible on the floor.

The cord was there to be pulled, and pulling it was rewarded.

AN OBSERVATION RECURRING ACROSS THE CHAPTER'S CASES.

CASE 155 INDUSTRIAL 1950S – PRESENT

Toyota Production System / Andon Cord

N Normalization of Deviance G Governance & Trust

IMPACT **Front-line authority to stop the line resolves most defects quickly at the source; defect-propagation cost minimized; the system adopted globally**

IN BRIEF

The andon cord lets any assembly-line worker signal a defect and — if it can't be resolved within the work cycle — stop Toyota's entire production line, handing the lowest-ranking person on the floor the power to halt operations worth millions per hour. The cord is trivially cheap; the authority it confers is the design. The case is decisive for capability engineering because when American automakers copied the cord in the 1980s and 1990s, workers were too afraid to pull it: the tool was present, the empowerment was not. Toyota's system works because it pairs the mechanism with a culture of psychological safety, fast supervisor response, no-blame problem-solving, and the codified "Five Whys" method. When the line stops at Toyota, the team treats it as a learning opportunity. The Andon Cord is the manufacturing twin of the Keystone nurse-authority intervention (Case 19).

BACKGROUND

In high-volume manufacturing, a defect that passes undetected propagates downstream, multiplying the cost of every later correction. Catching problems at the source requires the person who sees them — usually the lowest-ranking worker on the line — to be able to act, in an environment where stopping a line running at millions of dollars per hour is otherwise unthinkable. The economics and the authority structure point in opposite directions: the cheapest moment to fix a defect is the one at which the person who sees it has the least standing to halt production.[0]

THE INTERVENTION

As part of the Toyota Production System, Toyota installed the andon cord: a physical pull-cord that lets any worker signal a problem and, if unresolved, stop the entire line. The inversion of authority is the entire point — the cord itself costs almost nothing, while the protected authority it confers on a front-line worker is the actual design. Handing the lowest-ranking person the power to halt operations worth millions per hour deliberately resolves the contradiction the background poses: it puts the authority to stop exactly where the defect is first visible.[1]

HOW IT WORKED

The cord works because Toyota pairs the mechanism with a cultural system: psychological safety, a rapid supervisor-response protocol when the cord is pulled, no-blame root-cause analysis, and the codified "Five Whys" method. A stop is treated as a learning opportunity rather than a failure, so workers actually use it — the technical artifact and the protected

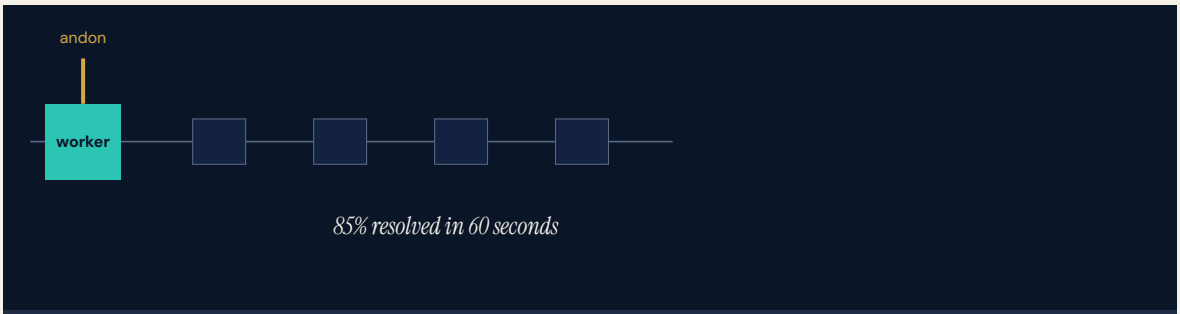
authority are inseparable. The rapid supervisor response is what makes the protection credible in practice: a worker who pulls the cord sees help arrive rather than blame, so the no-blame norm is demonstrated each time, not just asserted.[2]

THE EVIDENCE

The proof of the pairing is negative as well as positive. When American manufacturers copied the andon cord in the 1980s and 1990s, workers were too afraid to pull it; the artifact without the authority produced nothing. At Toyota, where the authority is protected, the great majority of activations are resolved within a minute and the system has been sustained and exported for decades. The natural experiment is unusually clean — the same physical cord produced opposite results across two settings, isolating the protected authority, not the hardware, as the variable that mattered.[3]

WHAT TRANSFERRED

The Andon Cord is the foundational evidence that authority interventions and technical artifacts are inseparable — the cord means nothing without the protected authority to pull it, and vice versa. It is the manufacturing counterpart of the Keystone nurse-stop authority (Case 19): same logic, different industry, same load-bearing element, and the same failure mode when only the artifact is copied. That the identical pattern recurs across manufacturing and medicine is what elevates it from a Toyota practice to a design principle: wherever the person who sees the problem lacks the standing to stop, copying the tool alone reproduces the failure.[4]



LOWEST-RANKING WORKER AUTHORIZED TO STOP THE LINE

> *When American manufacturers copied the andon cord, workers were too afraid to pull it.*
>
> PARAPHRASING LIKER, THE TOYOTA WAY (2ND ED., 2020)

REFERENCES

1. Liker, J. (2020), The Toyota Way (2nd ed.) — the andon cord, the cultural pairing, and the American imitation (paraphrased).
2. Spear, S. & Bowen, H. (1999), "Decoding the DNA of the Toyota Production System," HBR — the response protocol and embedded learning.
3. Shingo, S. (1989), A Study of the Toyota Production System — the technical mechanism.
4. Rother, M. (2009), Toyota Kata — the routines that sustain the practice.
5. Womack & Jones (1996), Lean Thinking — diffusion and the limits of surface imitation.

LE INSIGHT

The Andon Cord is the foundational evidence that authority interventions and technical artifacts are inseparable. The cord means nothing without the protected authority to pull it. The protected authority means nothing without the cord to act through. The pair is irreducible — and that irreducibility is the LENS co-optimization commitment in physical form.

LENS APPROACH

LENS uses the Andon Cord in LEN 2 as the foundational example of paired technical-cultural intervention, in LEN 8 to discuss cross-domain transfer (and why imitation without the cultural half fails), and in LEN 10 as a studio exemplar of minimal-artifact design.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
INDUSTRIAL →	N G →	D1 D2 **D3** D4 D5	3	4.1	3
		Human-System Collaboration			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Place the authority to stop exactly where the problem is first visible — with the front-line worker — resolving the contradiction that the cheapest moment to fix a defect is when the seer has the least standing.
- Pair the cheap artifact with the protected authority deliberately, recognizing the cord costs almost nothing while the authority it confers is the actual design.
- Stand up a rapid supervisor-response protocol so that pulling the cord brings help, not blame, making the no-blame norm credible from the first activation.

IN OPERATION / AFTER

- Sustain the pairing with no-blame root-cause analysis and a codified method (the Five Whys) so each stop becomes embedded learning rather than a one-off interruption.
- Demonstrate the protected authority repeatedly — help arriving within a minute — so the norm is shown each time rather than merely asserted in policy.
- When transferring the model, export the protected authority and response culture, not just the artifact, since copying the tool alone reproduces the failure mode.

REFLECTION QUESTIONS

1. Identify a technical artifact in your domain whose effectiveness depends entirely on a protected authority. Is the authority protected, or only declared?
2. American manufacturers copied the cord without the authority. What is the equivalent surface-level imitation you have observed in your domain, and what was missing?
3. Toyota's no-blame norm is demonstrated each time help arrives within a minute of a pull rather than blame. Design the visible, repeated response that would prove a protected authority is real in your domain rather than merely written into policy.

WHO BUILDS THIS safety science & organizational analysis · governance, ethics & measurement — plus domain experts and a learning engineer to integrate. TOOLS incident analysis · safety dashboards · governance frameworks · bias & evidence audits

CASE 157 TECHNOLOGY 2021 – PRESENT

AI-Augmented Coding Tools

[T] Training Gap [H] Human-System Interface

IMPACT **Tens of millions of developers using GitHub Copilot, Cursor, and peers; productivity gains documented; security and correctness implications still being characterized**

IN BRIEF

AI-augmented coding tools — GitHub Copilot, Cursor, Codeium, and peers — represent the largest real-time experiment in human-AI collaboration in this book, with tens of millions of developers using them daily. Published studies (Peng et al. 2023) document real short-term productivity gains; other work (Pearce et al. 2022) finds a substantial share of AI-generated completions in security-sensitive settings contain vulnerabilities, though a controlled study (Sandoval et al. 2023) found AI assistance did not significantly raise the rate of critical security bugs. The capability question is open: are developers becoming more capable, or more dependent? The short-term gains are real; the long-term consequences — especially for those who learn the craft with these tools always available — are not yet known. The discipline is being asked to define good before the longitudinal evidence is in.

THE SHIFT

AI coding assistants moved from novelty to infrastructure in a few years. GitHub Copilot, Cursor, Codeium, and similar tools now suggest, complete, and generate code for tens of millions of developers daily — the largest real-time experiment in human-AI collaboration in professional knowledge work to date, conducted not in a study design but in the live practice of an entire profession, with no control group and no agreed measure of what it is doing to the underlying craft.[0]

WHAT IS EMERGING

Two findings are accumulating in parallel. Controlled studies (Peng et al. 2023) document real short-term productivity gains. At the same time, the security picture is unsettled: Pearce et al. (2022) found a substantial fraction of Copilot completions in security-relevant scenarios contained vulnerabilities, while a controlled study by Sandoval et al. (2023) found AI assistance did not significantly increase the rate of critical security bugs. The two results do not cancel so much as mark how unsettled the picture is — output clearly rises, but the quality and safety of that output resist a single verdict.[1]

THE CAPABILITY QUESTION

The capability question is open and consequential: are developers using these tools becoming more capable, or more dependent? Short-term output rises, but whether the underlying skill grows or erodes — especially for those who learn the craft with the tools always present

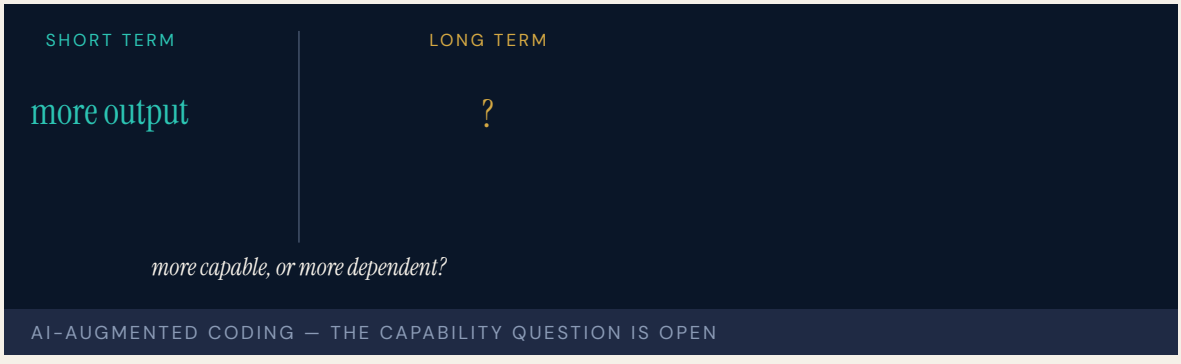
— is precisely what the productivity metrics cannot tell us, because a measure of how much code ships says nothing about whether the person shipping it could still produce or judge it without the assistant.[2]

EARLY EVIDENCE

The longitudinal evidence is not yet sufficient to answer. Dell'Acqua et al. (2023) found a "jagged frontier" in professional LLM use: performance improves on tasks inside the tool's competence and degrades on tasks just outside it, where users over-trust the output. The short-term gains are real; the long-term capability consequences remain uncharacterized — and the jagged frontier is hard to navigate precisely because its edge is invisible from inside the task, so the user cannot tell when they have crossed from where the tool helps to where it misleads.[3]

OPEN PROBLEMS

AI-augmented coding is the live frontier for human-AI teaming in professional knowledge work, and the discipline LENS represents is being asked to specify what good looks like before the long-term evidence is in. The open problem is the longitudinal study that could distinguish capability growth from capability erosion — and a training practice that keeps the human's skill on the growing side, built and adopted while a generation is already learning the craft with the tools always within reach. A July 2025 randomized controlled trial by METR sharpened the warning: experienced open-source developers allowed to use early-2025 AI tools took about 19% longer on real tasks, yet believed the tools had sped them up by roughly 20% — direct evidence that self-reported productivity can invert the true effect, exactly the measurement gap this case argues the discipline must instrument.[4]



AI-AUGMENTED CODING — THE CAPABILITY QUESTION IS OPEN

> *AI assistance changes what developers can do; it may also change what they need to know.*
>
> EDITORS' SYNTHESIS

REFERENCES

1. Peng et al. (2023), "The Impact of AI on Developer Productivity" — short-term productivity gains.
2. Pearce et al. (2022), "Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions."
3. Sandoval et al. (2023), "Lost at C," USENIX Security — AI assistance did not significantly increase critical security-bug rates.
4. Dell'Acqua et al. (2023), "Navigating the Jagged Technological Frontier" (HBS / BCG) — professional LLM use.
5. L. Bainbridge, "Ironies of Automation," Automatica 19(6) (1983) — the classic deskilling and over-reliance problem, applied to AI-augmented work.
6. METR (2025), "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity," arXiv:2507.09089 — the randomized trial finding a 19% slowdown against a self-reported speedup.

THE LEARNING ENGINEERING LENS

LE INSIGHT

AI-augmented coding is the live foundational case for human-AI teaming in professional knowledge work. The short-term gains are real. The long-term capability question — does the tool make the operator more capable, or more dependent — is the question the discipline must learn to ask and answer.

LENS APPROACH

The teaching point is a measurement-design problem, and it is the load-bearing one: productivity metrics count output and so cannot distinguish a developer whose skill is growing from one whose skill is quietly eroding under augmentation. The learning engineer's task is to build the instrument that separates those two. The design is a longitudinal, tool-removed probe — periodically measure each developer on representative tasks with the assistant withheld, scoring unaided correctness, debugging, and the ability to judge generated code, and track that aided-minus-unaided gap over time. A widening gap (rising aided output, flat-or-falling unaided competence) reads as skill-atrophy; a narrowing one reads as skill-growth. LENS uses this case in LEN 2 (human-AI teaming), in LEN 8 to build the capability-development instrument itself, and in LEN 10 (capstone) to have the student design the atrophy-versus-growth measurement for an augmented practice in their own domain.

DOMAIN	MODES AT STAKE	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
TECHNOLOGY →	T H →	D1 D2 **D3** D4 D5	6	5.2	3
		Human-System Collaboration			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Define what competence the human must retain independently of the tool, and design the workflow so that skill is exercised rather than quietly handed off.
- Engineer the assistant to surface the jagged-frontier edge — flagging where a task sits outside its reliable competence — so users do not over-trust output just beyond it.
- Keep verification of generated code, especially in security-relevant settings, a required step rather than an optional one, given the unsettled quality picture.

IN OPERATION / AFTER

- Run the longitudinal study that productivity metrics cannot substitute for, measuring whether underlying skill is growing or eroding over years of use.
- Monitor for over-reliance at the competence boundary, where the evidence shows performance degrading as users trust the tool past its reliable range.
- Track outcomes for practitioners who learned the craft with the tools always present, the cohort whose long-term capability is most uncertain.

REFLECTION QUESTIONS

1. In your domain, identify a class of practitioners whose work is currently being augmented by AI tools. What evidence would tell you whether their capability is growing or eroding?
2. Design the longitudinal study that would distinguish capability growth from capability erosion in an AI-augmented professional practice.
3. The "jagged frontier" is hard to navigate because its edge is invisible from inside the task. Design a signal or practice that would tell a practitioner in your domain when they have crossed from where the tool helps to where it misleads.

WHO BUILDS THIS learning science & instructional design · human factors & interaction design — plus domain experts and a learning engineer to integrate. TOOLS scenario simulation · competency models · usability testing · cognitive walkthroughs

Chapter 11

Disaster Prevention & Recovery — What Fails

Before the event, prevention regimes that thinned out; after it, response plans that met reality.

Every barrier had been inspected. What failed was the system deciding what counted as a barrier.

AN OBSERVATION RECURRING ACROSS THE CHAPTER'S CASES.

EDITOR'S NOTE · BEFORE AND AFTER

This part reads the corpus along a different axis than the domain parts that precede it: the lifecycle of a disaster. Every case here has two clocks. Before the event, there is a prevention regime — a certification, an inspection cadence, a defense-in-depth argument — whose quiet thinning is the real story. After the event, there is a response and a recovery, and the institutions that are (or are not) built so the same failure cannot recur. The chapter is therefore ordered in two arcs, marked below: prevention failures first, then response-and-recovery failures; and in the What Works chapter, prevention engineered as a deliverable, then recovery run as a discipline. Cases whose primary lessons belong to a domain — an aviation accident, a clinical harm — remain in their domain parts; what qualifies a case for this part is that its load-bearing lesson is the pre/post lifecycle itself.

BEFORE THE EVENT — PREVENTION

CASE 161 ENERGY INDUSTRIAL 2005

Texas City BP Refinery Explosion

[N] Normalization of Deviance [T] Training Gap [K] Knowledge & Institutional Memory [G] Governance & Trust

IMPACT 15 killed, 180 injured at BP's Texas City refinery; CSB found systemic safety-culture decay

IN BRIEF

On 23 March 2005 a startup at BP's Texas City refinery overfilled a distillation tower — operators were working from a miscalibrated level transmitter and a redundant high-level alarm that never sounded — and venting hydrocarbon vapor found an ignition source and exploded, killing 15 workers in trailers parked beside the unit and injuring 180. The Chemical Safety Board found accumulated safety-culture decay, deferred maintenance, and a celebrated cost-cutting program. Its central finding reshaped U.S. industrial safety: BP's **personal**-safety metrics were among the best in the industry, while its **process**-safety state — the integrity of the hazardous process itself — had been drifting, unmeasured. Texas City is the canonical evidence that the wrong measurement, reported as a win, is worse than no measurement at all.

BACKGROUND

BP's Texas City refinery, one of the largest in the U.S., had absorbed years of deferred maintenance and corporate cost-cutting, with instruments and alarms tolerated in a degraded state. On the surface it looked safe: its personal-injury rates were among the best in the industry.[0] The strong injury numbers actively reassured the organization, because the metric the company trusted most was the one that carried no information about the degrading instruments and alarms on which a safe startup actually depended.

WHAT HAPPENED

On 23 March 2005, during a unit startup, operators overfilled a distillation tower far past its safe level, working from a miscalibrated level transmitter that falsely showed the level declining while the tower overfilled, a redundant high-level alarm that never sounded, and an unreadable sight glass. Hydrocarbon vapor vented, drifted across the site, found an ignition source — an idling truck — and exploded. Fifteen workers in temporary trailers parked beside the unit were killed and about 180 injured.[1] Every contributor had been tolerated as routine — the faulty tower instruments, a high-level alarm reported broken repeatedly in the prior two years, the trailers parked beside a hazardous unit — so the startup was run blind to a danger the site had long since stopped seeing.

THE INVESTIGATION

The U.S. Chemical Safety Board's investigation became the most influential in the agency's history.[2] It found accumulated safety-culture decay, deferred maintenance, broken instruments tolerated as routine, trailers sited dangerously close to a hazardous unit, and cost

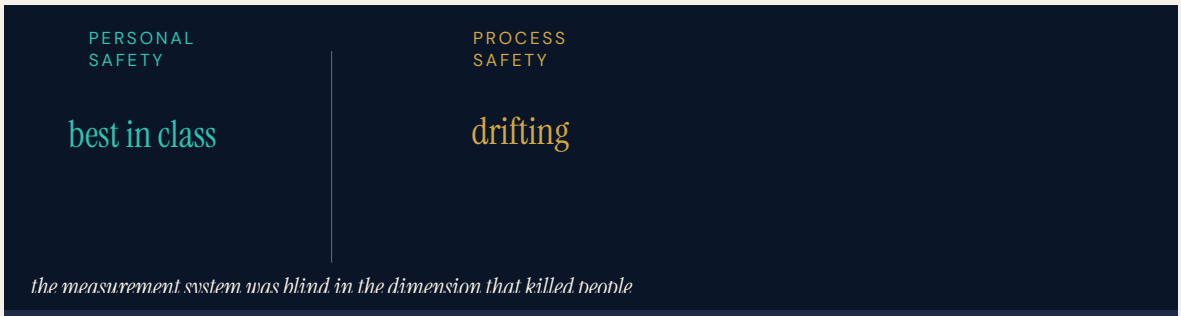
cutting — corporate 25% budget-reduction targets after the 1999 Amoco merger and again for 2005, and a 2003 site "1,000 Day Goals" program the CSB faulted for measuring personal safety and cost over process safety — celebrated as a success while it consumed the process-safety margin. The CSB drew the distinction that would reshape the field: "indicators of personal safety are not indicators of process safety."[3] The Board's force came from naming the deeper error as one of measurement: the company had not failed to measure but had measured the wrong dimension and then trusted the reassuring number it produced. BP's recordable personal-injury rate was excellent at the moment of the disaster, the front-page metric on which management-incentive plans paid out; the process-safety state of the unit that killed fifteen workers had no comparable reporting line at all.

THE CAPABILITY GAP

Texas City is the foundational evidence that an organization can measure the wrong thing and report excellent performance while its real capability gap widens. Personal-safety metrics — slips, trips, recordable injuries — carried no information about the integrity of the hazardous process, so the signal regime was blind in the very dimension that killed people. The wrong measurement, trusted, is worse than none.[4] A blank dashboard at least invites suspicion; a confident green reading on the wrong axis manufactures false assurance, which is why the excellent injury record made the process-safety drift harder to see rather than easier.

AFTERMATH & REFORM

The Baker Panel — chaired by former Secretary of State James Baker, commissioned by BP at the CSB's recommendation — reported in 2007 that the failure was a corporate one, not a Texas City one: a safety-culture decay extending across BP's five U.S. refineries, driven by cost pressure and management-of-change failures that the personal-safety metric system was structurally incapable of surfacing. BP invested heavily in process safety afterward, and the process-safety/personal-safety distinction — developed in the CCPS literature and codified in OSHA's 1992 process-safety-management standard — moved into mainstream U.S. industrial regulation after 2005, with API Recommended Practice 754 (2010) establishing process-safety leading and lagging indicators as an industry standard. The case's lasting contribution is a measurement lesson: count the thing that can kill you, not the thing that is easy to count.[5] That the distinction had been available in the CCPS literature and the OSHA standard before the explosion underscores the point: the knowledge existed, but the refinery's reporting had not been built to carry it upward where the hazard actually lived.



TEXAS CITY — THE WRONG MEASUREMENT, REPORTED AS A WIN

> *Indicators of personal safety are not indicators of process safety.*
>
> U.S. CHEMICAL SAFETY BOARD, BP TEXAS CITY FINAL INVESTIGATION REPORT, 2007

REFERENCES

1. U.S. Chemical Safety and Hazard Investigation Board, Refinery Explosion and Fire, Investigation Report 2005-04-I-TX (2007) — the startup, malfunctioning instruments, and 15 killed / 180 injured.
2. CSB (2007) — accumulated safety-culture decay, deferred maintenance, and the siting of occupied trailers beside a hazardous unit.
3. CSB (2007) — "indicators of personal safety are not indicators of process safety" (quoted).
4. Report of the BP U.S. Refineries Independent Safety Review Panel (the Baker Panel, 2007).
5. A. Hopkins, Failure to Learn: The BP Texas City Refinery Disaster (CCH, 2008).
6. OSHA Process Safety Management standard (29 CFR 1910.119, 1992) and the CCPS process-safety literature — the personal-vs-process-safety distinction.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Texas City is the foundational evidence that organizations can be measuring the wrong thing and reporting excellent performance while their actual capability gap widens. Personal-safety metrics had no information about process-safety state. The signal regime was blind in the dimension that killed people.

LENS APPROACH

Texas City is the canonical "measuring the wrong failure mode" case (induced 2.1; LENS D4/PT5), with cue/alert design (induced 3.1) and change-control (induced 5.4) as alternate anchors. LENS uses it in LEN 4 as the wrong-measurement case and in LEN 7 for the governance failure that allowed years of cost-cutting to be reported as wins. Studio projects design the process-safety measurement deliverable BP's executives should have been receiving in 2003. Adjacent to Wells Fargo (Case 148) at the measurement-system-blind-to-the-real-failure-mode layer.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
ENERGY INDUSTRIAL	N T K G	D1 D2 D3 D4 **D5** Navigating Sociotechnical Constraints	4	5.4	5

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Instrument process-safety state directly — barrier integrity, alarm health, instrument validity — rather than inferring safety from personal-injury rates.
- Treat degraded instruments and silent alarms as startup-blocking conditions, not routine items to defer past the next run.
- Set facility-siting rules that keep occupied trailers away from hazardous units as a design constraint, not a tolerated exception.

IN OPERATION / AFTER

- Audit whether the headline metric actually carries information about the hazard that can kill, and retire reassuring numbers that do not.
- Track deferred maintenance and tolerated-defect counts as process-safety leading indicators reported to executives.
- Verify that the process-vs-personal-safety distinction is wired into the reporting chain so it reaches the layer that funds maintenance.

REFLECTION QUESTIONS

1. Identify a "personal safety vs. process safety" equivalent in your domain. What capability gap is invisible to the metric your institution currently reports?
2. Design the process-safety dashboard that BP Texas City's executives should have been receiving in 2003.

WHO BUILDS THIS safety science & organizational analysis · learning science & instructional design · knowledge management & data engineering · governance, ethics & measurement — plus domain experts and a learning engineer to integrate. TOOLS incident analysis · safety dashboards · scenario simulation · competency models · knowledge bases · data pipelines · governance frameworks · bias & evidence audits

CASE 164 INDUSTRIAL 2017

Grenfell Tower

[G] Governance & Trust [T] Training Gap [K] Knowledge & Institutional Memory [N] Normalization of Deviance

IMPACT **72 killed in a residential tower fire in London; decades of regulatory failure**

IN BRIEF

On 14 June 2017 a fire spread up the exterior of Grenfell Tower, a London public-housing block, killing 72 people. It raced because the tower had been wrapped in combustible aluminium-composite cladding during a refurbishment. The public inquiry found the disaster the culmination of decades of failure: cladding firms engaged in "systematic dishonesty," marketing combustible materials as safe; regulators and inspectors missed effectively banned products across sixteen site visits; and the London Fire Brigade, whose "stay put" advice proved fatal, was unprepared for a cladding fire whose risks earlier incidents had already shown. The failure spanned manufacturers, regulators, inspectors, and responders — each contributing a piece, none owning the whole. Grenfell is the book's case for capability failure distributed across many hands.

BACKGROUND

Grenfell Tower was a 1970s public-housing block in West London, refurbished in 2015–16 with new exterior cladding intended to improve the building's appearance and efficiency. The cladding chosen used a combustible aluminium-composite material — installed despite safety experts' cautions that it was unsuitable for a high-rise, a warning that sat between the people who issued it and the people who specified the panels without ever stopping the decision.[0]

WHAT HAPPENED

On 14 June 2017 a kitchen fire broke out and, instead of staying contained as a tower's compartment design assumes, climbed the building's exterior on the combustible cladding, wrapping the tower in flame within minutes and defeating the very principle the building was meant to rely on. Residents, following long-standing "stay put" advice premised on that containment, remained in their flats as the route to safety closed around them; 72 people died.[1]

THE INVESTIGATION

The Grenfell Tower Inquiry found the fire the culmination of decades of failure by central government and every body responsible. Cladding companies had engaged in "systematic dishonesty," marketing combustible products as safe and corrupting the very test data buyers relied on; inspectors visited the site sixteen times and none noticed that effectively banned materials were in use, so sixteen chances to catch the hazard each passed it by.[2] The London Fire Brigade was unprepared: the risks of rapidly developing cladding fires were known from prior incidents — Knowsley Heights, Garnock Court, Shepherd's Court — but

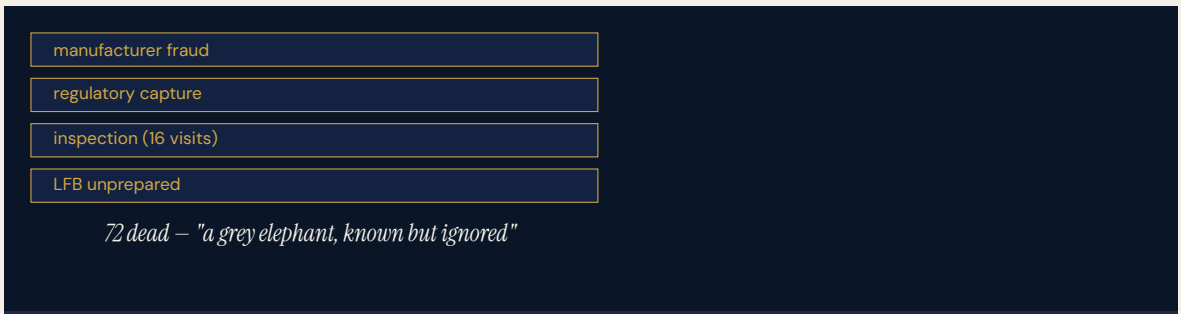
"this knowledge had not informed firefighting policies, practices or training," so each near-miss taught no one whose job was to act on it.[3]

THE CAPABILITY GAP

Grenfell is the book's strongest evidence that capability failure can be distributed across many actors, each contributing a small piece and none accountable for the whole. Manufacturer fraud, regulatory capture, inspection incompetence, training gaps, and lost institutional memory all converged on one building, and because each actor saw only its fragment, each could regard its own part as tolerable. It was what one might call a "grey elephant" — a danger known but ignored, hiding in plain sight — and the missing capability was anyone owning the integrated risk that everyone could see in part but no one held in full.[4]

AFTERMATH & REFORM

The inquiry's Phase 2 report (2024) and the government response (2025) drove an overhaul of building-safety regulation, cladding remediation, and fire-service doctrine — reforms that, between them, tried to assign the ownership the original system had left vacant.[5] Grenfell's lesson is the governance one this chapter turns on: when responsibility for a known risk is split across dozens of actors, the risk has, in effect, no owner — and a system with no owner for its gravest hazard will eventually pay for it, in the currency of the people living inside it.



LEVELS OF FAILURE: MANUFACTURE, REGULATION, INSPECTION, RESPONSE

> *This knowledge had not informed firefighting policies, practices or training.*
>
> GRENFELL TOWER INQUIRY, PHASE 1, 2019

REFERENCES

1. Grenfell Tower Inquiry, Phase 1 Report (2019) — the fire's spread up the cladding and the failure of "stay put."
2. Grenfell Tower Inquiry, Phase 2 Report (2024) — decades of failure and the combustible-cladding decision.
3. Phase 2 Report (2024) — cladding firms' "systematic dishonesty" and the inspection failures across sixteen visits.
4. Phase 1 Report (2019) — London Fire Brigade unpreparedness; "this knowledge had not informed firefighting policies, practices or training" (quoted).
5. UK Government response to the Grenfell Phase 2 report (2025) — building-safety and fire-service reform.
6. B. Hutter & M. Power (eds.), Organizational Encounters with Risk (2005) — distributed risk ownership.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Grenfell is the strongest evidence in the dataset that capability failure can be distributed across many actors, each of whom contributes a small piece, none of whom is accountable for the whole. The "grey elephant" framing — a known danger no one owns — describes a pattern that LENS treats as a primary governance problem in any high-consequence domain.

LENS APPROACH

LENS reads Grenfell through the institutional-memory-of-warnings channel (induced 7.4, with a 6.2 secondary): prior cladding-fire warnings existed — Lakanal House, the London Fire Brigade's own knowledge from earlier incidents — but none ever reached the refurbishment and cladding decision that could have acted on them. The capability deliverable is the channel that carries a known warning to the decision empowered to stop the work; the cladding firms' systematic dishonesty was a real aggravator but is the secondary thread, not the lesson.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
INDUSTRIAL →	G T K N →	D1 D2 D3 D4 **D5**	4	7.4	2 · 5
		Navigating Sociotechnical Constraints			

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Assign one accountable actor to own the building's integrated fire risk end to end, so no hazard can fall into the gaps between manufacturers, inspectors, and responders.
- Verify combustible-material claims against independent evidence rather than trusting vendor marketing, treating "systematic dishonesty" as a threat the process must defeat.
- Preserve the containment principle the building relies on: gate any cladding choice on whether it keeps a compartment fire from climbing the exterior.

IN OPERATION / AFTER

- Route every site inspection and near-miss into a shared record so sixteen visits cannot each miss the same banned material in isolation.
- Feed prior-incident knowledge into firefighting policy, training, and "stay put" doctrine, so lessons from earlier cladding fires actually change practice.
- Sustain a single line of accountability for the integrated hazard through the building's life, not only at refurbishment.

REFLECTION QUESTIONS

1. What is the "grey elephant" — the well-known risk that nobody owns — in your domain?
2. Design the deliverable that forces a single actor to own the integration risk that Grenfell distributed across dozens.
3. Sixteen inspections, prior fires, and expert cautions each touched a fragment of the Grenfell hazard, yet none assembled it. What mechanism in your domain could gather scattered partial warnings into one picture in front of someone empowered to stop the work?

WHO BUILDS THIS governance, ethics & measurement · learning science & instructional design · knowledge management & data engineering · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS governance frameworks · bias & evidence audits · scenario simulation · competency models · knowledge bases · data pipelines · incident analysis · safety dashboards

AFTER THE EVENT — RESPONSE AND RECOVERY

CASE 167 EMERGENCY MANAGEMENT DISASTER RESPONSE GOVERNMENT 2005

Hurricane Katrina — The FEMA/DHS Response Failure

G Governance & Trust K Knowledge & Institutional Memory

IMPACT Approximately 1,400 deaths under revised counts (earlier official estimates near 1,800; attribution varies by study) and roughly $125 billion in damage after Katrina's August 29, 2005 landfall; the most-investigated disaster response failure in U.S. history — House Select Bipartisan Committee "A Failure of Initiative" (February 2006), Senate HSGAC "Hurricane Katrina: A Nation Still Unprepared" (May 2006), and the White House Townsend report (February 2006) converged on a response that failed against a scenario the July 2004 Hurricane Pam exercise had simulated in detail thirteen months earlier; reform arrived as the Post-Katrina Emergency Management Reform Act of 2006

IN BRIEF

Hurricane Katrina made landfall on the Gulf Coast on 29 August 2005; the levee system protecting New Orleans failed, and the federal response failed with it. The casebook's spine for the case is that the missing capability was known before the event: in July 2004, the FEMA-funded Hurricane Pam exercise gathered roughly 300 local, state, and federal officials around a simulated slow-moving Category 3 hurricane whose storm surge overtopped the New Orleans levees, flooded the city, and stranded a population the plans could not move — almost exactly the scenario that arrived thirteen months later. Pam's action items were unfunded and unfinished when Katrina hit. In the event, FEMA's logistics system could not track or deliver commodities; the National Response Plan's Incident of National Significance machinery was invoked only after landfall; unified command across local, state, and federal levels never formed; communications interoperability collapsed; and the Superdome and Convention Center were known to television audiences before they were known to the federal operations picture. Death-toll estimates range from roughly 1,400 under revised counts to earlier official figures near 1,800, with attribution varying by study. Three major investigations followed, and the Post-Katrina Emergency Management Reform Act of 2006 rebuilt FEMA's authorities around a National Preparedness System.

BACKGROUND

New Orleans's vulnerability was not a discovery. The city sits largely below sea level behind a federal levee system, and the catastrophic-flood scenario had been anticipated for years in emergency-management planning. In July 2004 — thirteen months before Katrina — FEMA funded the Hurricane Pam exercise in Baton Rouge, bringing together roughly 300 local,

state, and federal emergency-response officials around a simulated slow-moving Category 3 hurricane with 120-mile-per-hour sustained winds, up to twenty inches of rain, and a storm surge that overtopped the levees in the New Orleans area. The scenario projected ten to twenty feet of flooding across most of the city and surrounding parishes, the destruction of hundreds of thousands of buildings, more than a hundred thousand residents unable to self-evacuate, and fatalities in the tens of thousands. Pam produced draft functional plans — search-and-rescue, sheltering, temporary medical care, debris — and a schedule of follow-on workshops to turn the drafts into executable, resourced capability. When Katrina made landfall, the follow-on work was partially unfunded and unfinished: the exercise had specified the capability requirement, and no institution had converted the specification into built capability.[0]

WHAT HAPPENED

Katrina made landfall on 29 August 2005; the levee system was breached rather than merely overtopped, and roughly eighty percent of New Orleans flooded. The response then discovered, in production, each capability the exercise had specified on paper. FEMA's logistics system could not track commodities in transit or confirm delivery — trucks of ice, water, and food moved without visibility into where they were or whether they arrived. The National Response Plan's machinery for a catastrophic event — the Incident of National Significance designation and the catastrophic-incident annex intended to push resources without waiting for state requests — was invoked only after landfall, in its first operational use, rather than driving pre-landfall mobilization. Unified command across city, state, and federal authorities never formed, and senior officials worked around the designed coordination structures. Communications interoperability collapsed as towers, switches, and emergency-operations centers flooded. Tens of thousands of residents concentrated at the Superdome and the Convention Center in deteriorating conditions; the Convention Center's situation was visible to television audiences before it entered the federal operations picture, and senior officials publicly disputed conditions that broadcast footage had already established. Estimates of the death toll range from roughly 1,400 under revised counts to earlier official figures near 1,800 — attribution of individual deaths to the storm varies by study — with damage of roughly $125 billion.[1]

THE INVESTIGATION

Three major investigations followed within a year, and their convergence is the case's evidentiary backbone. The House Select Bipartisan Committee's A Failure of Initiative (15 February 2006) found failures at every level of government, documented the Hurricane Pam exercise and its unconverted action items in detail, and framed the response in deliberate contrast to the 9/11 Commission's "failure of imagination": the imagination had been exercised, funded, and written down, and what failed was the initiative to act on it. The Senate Homeland Security and Governmental Affairs Committee's Hurricane Katrina: A Nation Still Unprepared (May 2006), built on more than 325 interviews and over 800,000 pages of documents, charged FEMA with failures across personnel deployment, communications, catastrophic planning, commodity pre-staging, medical and search-and-rescue deployment, and evacuation — and recommended dissolving FEMA and rebuilding its function inside DHS. The White House's The Federal Response to Hurricane Katrina: Lessons Learned (the Townsend report, 23 February 2006) issued 125 recommendations and 17

lessons, including the failure to maintain a common operating picture — the finding that formalized the television-before-operations problem.[2]

THE CAPABILITY GAP

Reform arrived through statute. The Post-Katrina Emergency Management Reform Act of 2006 (PKEMRA, Public Law 109-295, signed 4 October 2006) rebuilt FEMA's authorities inside DHS rather than dissolving it: it restored preparedness functions to the agency, set qualification requirements for the FEMA Administrator and gave the office a direct line to the President during catastrophes, authorized accelerated federal assistance in advance of a state request for precisely the catastrophic case Katrina represented, and mandated the national preparedness architecture — a national preparedness goal and system under which exercise findings are tracked as corrective actions rather than left as drafts. PKEMRA is the institutional codification of the case's lesson: it converted the discipline that Hurricane Pam lacked — a mechanism obligating someone to fund and finish what an exercise proves is missing — into statutory machinery. Whether the machinery holds is a separate question the casebook leaves open; what the statute establishes is that Congress located the failure where the investigations did, in the seam between demonstrated requirement and built capability.[3]

AFTERMATH & REFORM

The hedges the case carries are load-bearing. The death toll is contested: revised counts place it near 1,400, earlier official figures near 1,800, and the attribution of individual deaths — direct drowning versus indirect cardiovascular and displacement deaths — varies by study and jurisdiction; the casebook reports the range, not a number. The analytical spine is the Hurricane Pam seam: an exercised, documented capability requirement — levee overtopping, a flooded city, more than a hundred thousand residents who could not self-evacuate — that no institution converted into funded, staffed, rehearsed capability in the thirteen months before the event. This is the disaster-preparedness form of stated versus engineered requirements: the stated requirement existed in exercise documentation at a level of specificity the investigations found almost eerie, and the engineered capability did not exist at all. Response is the phase in which the gap is discovered in production, at the cost structure production imposes. The case asks the reader to account not for why no one foresaw the scenario — they did, in writing — but for why foresight without a funded owner is indistinguishable, on the day, from no foresight at all.

> *It remains difficult to understand how government could respond so ineffectively to a disaster that was anticipated for years, and for which specific dire warnings had been issued for days.*
>
> A FAILURE OF INITIATIVE, SELECT BIPARTISAN COMMITTEE, FEBRUARY 15, 2006.

REFERENCES

1. Select Bipartisan Committee to Investigate the Preparation for and Response to Hurricane Katrina (2006), A Failure of Initiative, H. Rept. 109-377, U.S. House of Representatives, February 15, 2006 — including the dedicated Hurricane Pam findings chapter.
2. U.S. Senate Committee on Homeland Security and Governmental Affairs (2006), Hurricane Katrina: A Nation Still Unprepared, S. Rept. 109-322, May 2006 — 325+ interviews, 838,000+ pages of documents reviewed.
4. U.S. Senate Committee on Homeland Security and Governmental Affairs (2006), Preparing for a Catastrophe: The Hurricane Pam Exercise, hearing, S. Hrg. 109-403, January 24, 2006 — testimony on the July 2004 exercise scope and unfinished follow-on work.
5. Post-Katrina Emergency Management Reform Act of 2006, Title VI of Public Law 109-295, signed October 4, 2006 — rebuilt FEMA authorities, accelerated pre-request federal assistance, national preparedness goal and system.

3. The White House (2006), The Federal Response to Hurricane Katrina: Lessons Learned (the Townsend report), February 23, 2006 — 17 lessons, 125 recommendations, including the common-operating-picture finding.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Katrina is the load-bearing case for a capability gap that was specified, exercised, and documented before the event and converted into built capability by no one. The Hurricane Pam exercise stated the requirement thirteen months in advance at near-exact fidelity; the response phase then discovered the unbuilt capability in production, at the cost structure production imposes. Foresight without a funded owner is indistinguishable, on the day, from no foresight at all.

LENS APPROACH

Katrina is the stated-versus-engineered-requirements case at national scale (induced 1.1; LENS D1/PT1; LEO-1). The induced anchor is 1.1 — engineered versus stated requirements — because the case's spine is precisely that seam: Hurricane Pam produced a stated requirement of unusual fidelity, and nothing engineered it into funded, staffed, rehearsed capability before the event; the response failures are downstream expressions of the unconverted requirement, not independent causes. LENS uses it in Domain 1 (Systems Analysis) for the discipline of tracing a documented requirement to the built capability that does or does not exist behind it, with the pre/post disaster lifecycle of this part as the organizing frame. Pair with the prevention-thinning cases earlier in the chapter and with the PKEMRA reform arc for the institution-building sequel.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
EMERGENCY MANAGEMENT DISASTER RESPONSE GOVERNMENT	G K	D1 D2 D3 D4 D5 Systems Analysis	1	1.1	1

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Treat an exercise finding as an open capability requirement with a named owner, a funding line, and a completion date; Hurricane Pam's draft plans and follow-on workshop schedule specified the requirement, and the case's load-bearing observation is that specification without a funded owner produced no capability in thirteen months.
- Verify the catastrophic-response machinery by operating it before the catastrophe; the Incident of National Significance designation and the push-resources-without-request annex had never been exercised end-to-end in anger, and their first operational use came after landfall of the event they were designed for.
- Build the common operating picture as an engineered deliverable, not an aspiration; the test is concrete — the operations center should know what broadcast television knows, and a response organization that fails that test has a situational-awareness capability gap, not a communications inconvenience.

IN OPERATION / AFTER

- Convert investigation findings into statutory obligation where institutional memory alone will not hold; PKEMRA's corrective-action and national-preparedness mandates are the codified form of the discipline Pam lacked, and the case teaches the statute as the institutional response, not as the proof the gap is closed.
- Preserve the contested death-toll attribution in any teaching use; the range and its methodological instability are themselves evidence about how poorly instrumented the response was, and smoothing to a single number discards that signal.
- Pair the case in syllabi with the prevention-side cases in this part so the pre/post lifecycle reads as one arc: a prevention regime that thinned, a rehearsed requirement that went unbuilt, and a response phase that discovered the unbuilt capability in production.

REFLECTION QUESTIONS

1. Identify an exercise, red-team, or audit finding in your organization that specified a missing capability and was closed as "documented" rather than "built." What would a funded owner, a completion date, and a re-verification event have looked like, and what converts a finding into an obligation in your setting?

2. The Incident of National Significance machinery saw its first operational use during the event it was designed for. Pick a rarely-invoked escalation mechanism in your domain and specify the rehearsal that would verify it end-to-end — including the authority handoffs — before it is needed in production.

3. The Convention Center was visible to television audiences before it entered the federal operations picture. What is the equivalent external signal in your domain that would outrun your operations picture, and what engineered channel would make your institution aware of what the public already knows?

WHO BUILDS THIS governance, ethics & measurement · knowledge management & data engineering — plus domain experts and a learning engineer to integrate. TOOLS governance frameworks · bias & evidence audits · knowledge bases · data pipelines

Chapter 12

Disaster Prevention & Recovery — What Works — and the Frontier

Prevention as a deliverable before the event; recovery as a discipline after it.

The disaster that did not happen had a budget, an owner, and a deadline.

AN OBSERVATION RECURRING ACROSS THE CHAPTER'S CASES.

BEFORE THE EVENT — PREVENTION

CASE 173 NAVAL ENGINEERING DEFENSE SAFETY CERT 1963 – PRESENT

Navy SUBSAFE — Requirements as a Non-Negotiable Sustainment Deliverable

G Governance & Trust K Knowledge & Institutional Memory H Human-System Interface

IMPACT **From 1915 to 1963 the US Navy lost 16 submarines to non-combat causes; since 1963 it has lost one — USS Scorpion — which was not SUBSAFE-certified. The Columbia Accident Investigation Board cited SUBSAFE as a model NASA could emulate**

IN BRIEF

USS Thresher was lost with 129 aboard on April 10, 1963. Within fifty-four days the US Navy created SUBSAFE — a program that certifies design, material, fabrication, and testing for every component inside the submarine's watertight-integrity boundary and its safe-recovery systems. The requirements were issued by December 20 of that same year. The program demands what it calls "Objective Quality Evidence" for every step — verifiable fact, not probabilistic assessment — and pairs that with annual training and recurring audits across the entire fleet lifecycle. The documented result is a step-change in non-combat submarine loss rates: 16 losses across the 48 years before SUBSAFE; one loss (USS Scorpion, not SUBSAFE-certified) across the 63 years since. The Columbia Accident Investigation Board cited SUBSAFE in 2003 as one of three programs that could be models for NASA. The honest hedge survives: the zero-loss record is correlational across decades with many co-varying factors — submarine design, reactor maturity, operating procedures, intelligence environment — and SUBSAFE's own program literature notes that the requirements can look "excessive." The case is the archetype of treating capability requirements as a recurring, auditable, non-waiverable deliverable across the entire system lifecycle, with the hedges that decades-long capability-engineering claims have to carry.

BACKGROUND

USS Thresher, the lead boat of a new class of US nuclear attack submarines, was lost on April 10, 1963 during post-overhaul sea trials off Cape Cod. All 129 aboard died. The investigation identified a likely sequence — silver-brazed piping joint failure, flooding, electrical fault that scrammed the reactor, inability to blow ballast tanks fast enough to recover — that pointed not to a single defective component but to a gap in the whole way the fleet certified the watertight-integrity boundary and the systems for recovering from a flooding casualty.[o]

THE INTERVENTION

The Navy's response was institutional speed of an unusual order. SUBSAFE was created within fifty-four days of the loss; formal requirements were issued by December 20, 1963. The program scopes itself to two things: the watertight-integrity boundary (every component that holds back seawater at depth), and the safe-recovery systems (the ballast-blow

and propulsion systems that get a boat to the surface if flooding starts). Inside that scope it demands certification of design, material, fabrication, testing, and configuration control — with what the program calls "Objective Quality Evidence" attached at every step. That phrase is the load-bearing cultural artifact of the program: verifiable fact, not probabilistic assessment, is what the certification rests on.[1]

HOW IT WORKED

What makes SUBSAFE a sustainment intervention rather than a design intervention is the lifecycle discipline. Annual training across the fleet, recurring audits at shipyards and tenders, change-control on every modification, and a culture that treats requirements as non-waiverable artifacts the program-management chain cannot trade away under schedule pressure. The certification is not done at launch; it is the condition of being allowed to dive. Each overhaul re-engages the certification process. The cost is real — the program's own histories concede the requirements can look "excessive" — and the program treats that cost as the price of the capability the certification produces.[2]

THE EVIDENCE

The documented outcome is one of the cleanest before/after records in the safety-engineering literature. From 1915 to 1963 the Navy lost 16 submarines to non-combat causes (collision, flooding, equipment failure, fire). Since 1963 it has lost one — USS Scorpion in 1968, the only post-1963 loss that was not SUBSAFE-certified. In 2003 the Columbia Accident Investigation Board, examining the loss of the Space Shuttle Columbia, cited SUBSAFE in its final report as the capability-certification model NASA should adopt for human spaceflight. The endorsement is from an investigation body with no Navy institutional stake, examining a different catastrophic-system domain.[3]

WHAT TRANSFERRED

The hedge has to survive into the case. The zero-loss record since 1963 is correlational across more than six decades and many co-varying factors: submarine design generations, reactor maturity, training systems, operating procedures, intelligence environment, the absence of certain operational stressors that the Cold War sometimes produced. Attributing the entire outcome to SUBSAFE alone overstates what the evidence can support. What the evidence does support is that the **program** — the requirements discipline, the Objective Quality Evidence standard, the lifecycle audit cycle, the non-waiverable culture — has been a defining feature of the capability since 1963, and has survived endorsement by an independent investigation in a different domain. The case teaches the requirements-as-sustainment-deliverable form at its strongest, with the honest hedge that decades-long capability claims have to carry.[4]

> *Objective Quality Evidence — verifiable fact, not probabilistic assessment — is the cultural artifact the program is built on.*
>
> EDITORS' SYNTHESIS OF THE SUBSAFE PROGRAM LITERATURE.

REFERENCES

1. Rear Admiral Paul E. Sullivan, statement to House Science Committee (2003), NASA/Columbia archive — primary congressional record on SUBSAFE.
4. Columbia Accident Investigation Board (2003), final report — Volume I, endorsement of SUBSAFE as a model for NASA.
5. US Navy NAVSEA, SUBSAFE program documentation — operating program publications.

2. MIT Press, "SUBSAFE: An Example of a Successful Safety Program" — book chapter (open access).
3. NASA SMA (2006), "USS Thresher Lessons Learned" briefing — safety-message archive.

THE LEARNING ENGINEERING LENS

LE INSIGHT

SUBSAFE is the archetype of treating capability requirements as a recurring, auditable, non-waiverable deliverable across the entire system lifecycle. The before/after non-combat-loss record is one of the cleanest in safety engineering — and correlational across many co-varying factors over six decades. The hedge is part of the case.

LENS APPROACH

SUBSAFE is the canonical sustainment-engineering case (induced 1.4; LENS D1/PT3). LENS uses it in Domain 1 (Systems Analysis) for the requirements-as-deliverable discipline; in Domain 4 (Test and Evaluation) for the Objective-Quality-Evidence standard and the recurring-audit cycle; and in Domain 5 (Navigating Sociotechnical Constraints) for the non-waiverable culture that resists schedule pressure. Adjacent to the nurse-ratios case (Case 11) at the requirements-becomes- engineered layer, and to the WHO Surgical Checklist (Case 23) at the mandatory-mechanism layer.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
NAVAL ENGINEERING DEFENSE SAFETY CERT	G K H	D1 D2 D3 D4 D5 Systems Analysis	3	1.4	1 · 4

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Scope the certification boundary tightly — for SUBSAFE, the watertight-integrity boundary and the safe-recovery systems — so the discipline is enforceable, not aspirational across an undifferentiated whole.
- Make "Objective Quality Evidence" the cultural standard: verifiable fact at every certification step, not probabilistic assessment, and not signature-without-evidence.
- Treat the requirements as non-waiverable artifacts the program-management chain cannot trade away under schedule pressure; the program's resistance to waivers is the program.

IN OPERATION / AFTER

- Operate the lifecycle discipline as the program: annual training, recurring audits, change-control on every modification, certification re-engaged at each overhaul.
- Carry the correlational hedge in any communication of the outcome record; a decades-long zero-loss record across co-varying factors is the strongest available evidence the program works, not closed proof.
- Treat external endorsement (Columbia Accident Investigation Board) as a teaching artifact about the **form** of the program, transferable to other catastrophic-system domains under their own scope discipline.

REFLECTION QUESTIONS

1. Identify a capability in your domain where the certification is done at launch and not re-engaged across the lifecycle. What is the analog of SUBSAFE's annual training and recurring audit — and what is the resistance to it?

2. Specify what "Objective Quality Evidence" would mean in your context: verifiable fact at every step rather than signature-without-evidence. Which steps in your current process would not survive that standard?

3. SUBSAFE's outcome record is correlational across many co-varying factors. What is the minimum additional evidence you would require before treating a similar long-run record in your domain as evidence the program is what produced the outcome?

WHO BUILDS THIS governance, ethics & measurement · knowledge management & data engineering · human factors & interaction design — plus domain experts and a learning engineer to integrate. TOOLS governance frameworks · bias & evidence audits · knowledge bases · data pipelines · usability testing · cognitive walkthroughs

AFTER THE EVENT — RESPONSE AND RECOVERY

CASE 176 HEALTHCARE INDUSTRIAL 1982

Tylenol Recall

G Governance & Trust N Normalization of Deviance

IMPACT **Foundational U.S. corporate crisis-management case; produced tamper-evident packaging regulation and modern recall practice**

IN BRIEF

In 1982, seven people in the Chicago area died after taking Tylenol capsules laced with potassium cyanide. Not knowing who was responsible or how widespread the tampering was, Johnson & Johnson recalled every Tylenol product nationwide — 31 million bottles, at a cost of roughly $100 million — suspended advertising, and engaged openly with the FBI and FDA. The response was unprecedented in U.S. corporate practice, and it was a direct application of the J&J Credo, written in 1943, which had pre-specified that the company's first responsibility was to its customers. The reform that followed reshaped consumer packaging worldwide — tamper-evident seals and blister packs — and the FDA issued tamper-resistant-packaging rules within months. Tylenol recovered its market share within a year.

BACKGROUND

Johnson & Johnson's corporate Credo, written in 1943, pre-specified that the company's first responsibility was to the patients and consumers who used its products, ahead of shareholders. For four decades it was a statement of values; in 1982 it became an operational decision rule under extreme pressure. The ordering was explicit — customers ahead of shareholders — which is precisely the ranking that crisis pressure inverts, so committing to it in advance pre-decided the hardest trade-off before it had to be faced.[0]

THE INTERVENTION

After seven people in the Chicago area died from Tylenol capsules laced with potassium cyanide, and with the source and scope of the tampering unknown, Johnson & Johnson recalled every Tylenol product nationwide — about 31 million bottles, at a cost near $100 million — suspended all advertising, and engaged publicly with the FBI and FDA rather than minimizing exposure. Recalling nationwide despite the tampering being known only in Chicago was the decisive choice — it treated the unknown scope as a reason to protect every customer rather than as room to limit the company's own exposure.[1]

HOW IT WORKED

The load-bearing element was a commitment pre-committed in writing. Because the Credo had already decided, decades earlier, that the customer came first, the 1982 leadership did not have to improvise an ethical calculus under crisis — it executed a pre-made decision. CEO James Burke later credited the Credo with making clear "exactly what we were all about" the moment the deaths occurred. Pre-commitment worked because it moved the decision

out of the moment of maximum pressure — when fear and legal caution push hardest toward minimization — and into a calmer time when the right ordering could be set down without that distortion.[2]

THE EVIDENCE

The response, unprecedented in U.S. corporate practice, preserved public trust: Tylenol recovered its market share within a year despite the enormous short-term cost. The case became the canonical positive example in business education of crisis response driven by capability commitment rather than legal minimization. The market recovery is what makes the case persuasive rather than merely admirable — the $100 million spent protecting customers was repaid in the trust that brought them back, so the pre-committed choice proved sound on its own terms.[3]

WHAT TRANSFERRED

The reform reshaped consumer-product packaging worldwide — tamper-evident seals, blister packs, and caplet forms became standard — and the FDA promulgated tamper-resistant-packaging regulations within months. The deeper transfer is the principle that values must be pre-committed in writing to be operational under stress, not invented in the moment. The packaging reform and the decision-rule principle are the two layers of the transfer — one a physical safeguard against the specific threat, the other an institutional safeguard against the improvisation that crisis invites.[4]

TYLENOL — VALUES PRE-COMMITTED IN WRITING, EXECUTED UNDER CRISIS

> *The Credo is all about the consumer. When those seven deaths occurred, the Credo made it very clear at that point exactly what we were all about.*
>
> JAMES BURKE (JOHNSON & JOHNSON CEO) — CLOSING SENTENCE QUOTED IN LASTING LEADERSHIP (WHARTON); OPENING LEAD IS THE EDITORS' PARAPHRASE

REFERENCES

1. Kaplan, T. (2014), The Tylenol Crisis — the recall and corporate response.
2. James Burke (J&J CEO), in Lasting Leadership (Wharton) — the Credo quote.
3. Greyser, S., Johnson & Johnson: The Tylenol Tragedy (HBS case, 1992) — market recovery and crisis management.
4. FDA Final Rule on Tamper-Resistant Packaging (1982) — the packaging reform.
5. Edmondson, A. (2018), The Fearless Organization — pre-committed values under stress.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Tylenol is the canonical positive case for institutional response to crisis. The capability that was load-bearing was the pre-specified institutional commitment in the Credo. The crisis decision had been made decades earlier; in 1982 it was executed. The case is the strongest evidence in the business-ethics dataset that values must be pre-committed in writing to be operational under stress.

LENS APPROACH

LENS uses Tylenol in LEN 7 as the foundational positive case for institutional governance under crisis and in LEN 10 (capstone) as a worked example of pre-committed capability that executed under operational pressure.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
HEALTHCARE INDUSTRIAL	G N	D1 D2 D3 D4 **D5** Navigating Sociotechnical Constraints	3	4.4	5

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Pre-commit the hardest trade-off in writing before the crisis — rank customer safety ahead of shareholder exposure — so leadership executes a pre-made decision rather than improvising under pressure.
- Set the rule in a calm period when fear and legal caution cannot distort the ordering, since those forces push hardest exactly when the decision must be made.
- Make the commitment concrete enough to act on — a nationwide recall, open engagement with regulators — so unknown scope becomes a reason to protect everyone rather than room to limit exposure.

IN OPERATION / AFTER

- Pair the institutional decision rule with a physical safeguard against the specific threat (tamper-evident packaging) so the response addresses both the improvisation problem and the vulnerability.
- Treat the preserved trust and market recovery as the measure that the pre-committed choice was sound, not merely admirable, and document it to defend the principle internally.
- Embed the commitment durably enough that it survives leadership turnover, so the next crisis meets the same pre-decided rule rather than a fresh improvisation.

REFLECTION QUESTIONS

1. What is your institution's equivalent of the J&J Credo, and is it operational under crisis or aspirational?
2. Pre-commitment is hard to enforce later. Design the institutional architecture that makes a Tylenol-style response the only available option in the worst case.
3. J&J recalled nationwide while the tampering was known only in Chicago, treating unknown scope as a reason to protect everyone. Identify a decision in your domain where uncertainty currently licenses minimizing exposure, and write the pre-committed rule that would flip it toward protection instead.

WHO BUILDS THIS governance, ethics & measurement · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS governance frameworks · bias & evidence audits · incident analysis · safety dashboards

CASE 179 PUBLIC HEALTH PANDEMIC RESPONSE DATA INFRASTRUCTURE 2020 – 2023

The Johns Hopkins COVID-19 Dashboard — Evidence Infrastructure at Decision Speed

G Governance & Trust K Knowledge & Institutional Memory

DISCLOSURE Contribution: an editor leads the Johns Hopkins APL group that contributed engineering support to this system; the editor was not personally involved in the build. Included on the published peer-reviewed record and independent recognition.

IMPACT Launched 22 January 2020 by two researchers and became the world's de facto epidemiological reference for the COVID-19 pandemic: peak single-day usage above 4.6 billion data requests (29 March 2020), over 226 billion feature-layer requests and 3.6 billion page views by June 2022 per the peer-reviewed record; governments, newsrooms, and researchers built directly on its openly published data; 2022 Lasker-Bloomberg Public Service Award; ceased data collection 10 March 2023

IN BRIEF

On 22 January 2020 — one day after the United States reported its first COVID-19 case — Lauren Gardner's Center for Systems Science and Engineering (CSSE) at the Johns Hopkins Whiting School, with graduate student Ensheng Dong, launched a public dashboard tracking reported COVID-19 cases in near-real time (Dong, Du & Gardner, The Lancet Infectious Diseases, published online February 2020). What began as a manually curated map became within weeks the world's de facto epidemiological reference: billions of data requests per day at peak, an underlying GitHub repository that by press accounts became one of the most-forked data repositories on the platform, and governments, newsrooms, and researchers building directly on its openly published data. The engineering story is the case: manual curation was replaced by a semi-automated pipeline with anomaly detection; the Johns Hopkins Applied Physics Laboratory (APL) partnered with CSSE on pipeline scaling, validation, and anomaly-detection engineering; and the university stood up the Coronavirus Resource Center (CRC, March 2020), joining medicine, public health, and engineering to add testing, hospitalization, vaccine, and equity data layers. The load-bearing hedge is preserved throughout: the dashboard is simultaneously a success case and evidence of the system gap it filled — for much of 2020 no government source operated at its speed or completeness, and its sustainment was never structural. Independent recognition: the 2022 Lasker-Bloomberg Public Service Award to Gardner. The CRC ceased data collection on 10 March 2023. The COI render under the title (an editor leads an APL group that contributed engineering support) is binding.

BACKGROUND

When a novel pathogen emerges, the first institutional casualty is measurement: case counts scattered across national health ministries, provincial bulletins, and press briefings, in inconsistent formats, on inconsistent cadences, with no single place a decision-maker could look to see the outbreak as it unfolded. On 22 January 2020, a day after the first reported US case, Gardner and Dong at JHU CSSE launched what the team's own peer-reviewed retrospective calls the first global real-time coronavirus surveillance system: an interactive web dashboard, initially fed by hand from a shared spreadsheet, publishing every number it displayed as freely available open data.[0]

THE INTERVENTION

The engineering trajectory is the case's first contribution. Manual curation could not survive the pandemic's growth curve, and the team replaced it deliberately: trusted-source identification and validation, autonomous web-scraping collection, a curation layer built around an in-house anomaly-detection service that flagged variations exceeding configured thresholds for human review before publication, and open data sharing through GitHub and hosted feature layers. The Johns Hopkins Applied Physics Laboratory partnered with CSSE on this build-out — the published record credits APL with engineering support to pipeline scaling, near-real-time processing, and anomaly detection, with APL data scientists among the reviewers of flagged discrepancies. The editors record here, with the adjacent conflict disclosed under the title, their direct knowledge that the Laboratory's engineering role in this transition was substantially larger than the public record conveys — closer to being responsible for the production pipeline than "support." That assessment is editorial testimony, not a published finding; the citations carry only what is published. The system held under extraordinary load: a peak of more than 4.6 billion requests in a single day (29 March 2020), and over 226 billion feature-layer requests by June 2022.[1]

HOW IT WORKED

The institutional trajectory is the second contribution. In March 2020 the university stood up the Coronavirus Resource Center, joining the engineering school's dashboard with the schools of medicine and public health to add the data layers the raw case map could not carry: testing rates, hospitalization, vaccination, and — through its racial-data- transparency work — a public accounting of which US states did and did not report race- and ethnicity-stratified data, turning the absence of equity data into a visible, trackable fact rather than a silent gap.[2]

THE EVIDENCE

The capability gap the dashboard exposed is as load-bearing as the artifact itself. For much of 2020, a professor, her graduate students, and a volunteer-plus-institutional support network were faster and more complete than any government source; the world's newsrooms and health agencies cited a university web map because no national public-health data infrastructure operated at decision speed. The 2022 Lasker-Bloomberg Public Service Award to Gardner is the independent recognition of the intervention — and, read carefully, an indictment of the vacuum that made a two-person launch the global system of record.[3]

WHAT TRANSFERRED

The ending preserves the hedge. The CRC ceased data collection on 10 March 2023, as states slowed their reporting cadences and federal tracking improved enough to warrant the exit; the full data record from 22 January 2020 to 10 March 2023 remains freely available. The team's own retrospectives name the fragility: heroic effort — students, volunteers, and institutional partners substituting for public infrastructure — was never a structural solution, and the dashboard's success does not demonstrate that the next pandemic will find its equivalent standing ready. The case is a success case and evidence of the system gap it filled; both readings are binding.

> *All data collected and displayed are made freely available, initially through Google Sheets and now through a GitHub repository, along with the feature layers of the dashboard.*
>
> DONG, DU & GARDNER (2020), THE LANCET INFECTIOUS DISEASES.

REFERENCES

1. Dong, Du, & Gardner (2020), "An interactive web-based dashboard to track COVID-19 in real time," The Lancet Infectious Diseases 20(5), 533 – 534, doi:10.1016/S1473-3099(20)30120-1.
2. Dong et al. (2022), "The Johns Hopkins University Center for Systems Science and Engineering COVID-19 Dashboard: data collection process, challenges faced, and lessons learned," The Lancet Infectious Diseases 22(12), e370 – e376, doi:10.1016/S1473-3099(22)00434-0.
3. Johns Hopkins University Hub (10 March 2023), "Johns Hopkins COVID-19 data hub ends after three years" — Coronavirus Resource Center closure announcement.
4. Lasker Foundation (2022), Lasker-Bloomberg Public Service Award citation — "The COVID-19 Dashboard" (Lauren Gardner); companion award essay in JAMA.
5. NPR (10 February 2023), "As the pandemic ebbs, an influential COVID tracker shuts down."

THE LEARNING ENGINEERING LENS

LE INSIGHT

The JHU COVID-19 dashboard is the evidence-infrastructure intervention case: a two-person launch that became the world's epidemiological reference because it measured the thing everyone needed measured, published the data openly, and engineered its way off manual curation before scale broke it. The binding hedge: it is simultaneously a success case and proof of the public-infrastructure gap it filled — sustainment was never structural, and the capability wound down in March 2023 with the gap's permanent closure still an open question. COI under the title (editor leads an APL group that contributed engineering support) is binding.

LENS APPROACH

The JHU COVID-19 dashboard is the evidence-infrastructure-at- decision-speed case (induced 2.1; LENS D4/PT2). LENS uses it in Domain 4 (Test and Evaluation) for building the measurement the institution actually needs — trusted sourcing, anomaly detection, human review, open publication — under real-time load; in Domain 1 (Systems Analysis) for reading the intervention as a diagnosis of the missing national data infrastructure; and in Domain 5 (Navigating Sociotechnical Constraints) for the racial-data-transparency layer that made absent equity data a public, trackable fact. Pair with failure-mode cases where institutions measured what was convenient rather than what mattered. COI render is binding.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
PUBLIC HEALTH PANDEMIC RESPONSE DATA INFRASTRUCTURE	G K	D1 D2 D3 **D4** D5 Test and Evaluation	2	2.1	4

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Publish the evidence base itself, not just the visualization — open data with transparent sourcing let governments, newsrooms, and researchers build on the pipeline instead of duplicating it.
- Engineer the transition off heroics early: replace manual curation with automated collection plus anomaly detection and human review before the growth curve makes the manual process fail publicly.
- Instrument the gaps as deliberately as the counts — tracking which states reported race/ethnicity-stratified data made missing equity data a visible fact that could be pushed on, not a silent omission.

IN OPERATION / AFTER

- Name the vacuum in the retrospective: the peer-reviewed lessons-learned record states plainly what the system substituted for, so the success cannot be read as evidence the gap is closed.
- Archive the full data record openly at shutdown so the three-year evidence base remains auditable and reusable after collection ends.
- Convert the demonstrated capability into a structural requirement — the case's unresolved question is whether public-health data infrastructure at decision speed gets built and funded as infrastructure, or waits for the next volunteer build.
- Commission the first-person account: the APL-CSSE production-pipeline build is a natural Practice Flywheel narrative for a future edition — the inside view of an iteration the public record undersells, told by the people who ran it and published under the same disclosure that governs this case.

REFLECTION QUESTIONS

1. Identify the measurement your institution would need on day one of a fast-moving crisis. Does it exist at decision speed today, or would a volunteer build have to substitute? What would the JHU-equivalent stopgap look like in your domain — and who would staff it?

2. The dashboard's pipeline paired automation with an anomaly-detection-plus-human-review gate before publication. Specify that gate for a data product you own: what thresholds flag an anomaly, who reviews, and what is the cost of publishing a wrong number at global scale?

3. The CRC tracked which states failed to report race/ethnicity-stratified data, making the gap itself a published metric. Choose a dataset in your domain with a known equity gap: what would it take to measure and publish the gap's extent, rather than footnoting its existence?

WHO BUILDS THIS governance, ethics & measurement · knowledge management & data engineering — plus domain experts and a learning engineer to integrate. TOOLS governance frameworks · bias & evidence audits · knowledge bases · data pipelines

Chapter 13

Algorithms, Governance & Public Systems — What Fails

When delegation to systems outruns accountability for them.

No one could say who had decided, because no one had.

AN OBSERVATION RECURRING ACROSS THE CHAPTER'S CASES.

CASE 184 TECHNOLOGY GOVERNMENT 1999 – 2015

UK Post Office Horizon Scandal

G Governance & Trust H Human-System Interface K Knowledge & Institutional Memory

IMPACT more than 900 sub-postmasters wrongfully convicted; 236 imprisoned; at least 13 linked suicides; described as the most widespread miscarriage of justice in UK history

IN BRIEF

The UK Post Office's Horizon accounting system, built by Fujitsu and rolled out in 1999, generated phantom shortfalls in sub-postmasters' branch ledgers. Rather than accept the software was at fault, the Post Office prosecuted them for theft and false accounting — about 700 cases itself, its Horizon evidence driving roughly 280 more prosecutions by other bodies, more than 900 wrongful convictions in all; people were imprisoned, bankrupted, and driven to suicide, in what is now called the most widespread miscarriage of justice in UK history. Internal documents showed engineers had known about Horizon bugs throughout. Convictions began to be quashed in December 2020, and a public inquiry continues. The failure ran through the prosecutor and the courts: each accepted "the computer said so" as authoritative because no actor had the standing or expertise to challenge it. Horizon is the book's case for institutional deference to a flawed algorithm.

BACKGROUND

The UK Post Office ran thousands of branches through sub-postmasters — local operators personally liable for any shortfall in their accounts, a liability that put each operator's livelihood behind the numbers the system reported. In 1999 it deployed Horizon, an accounting system built by Fujitsu, to track every branch's ledger, making the software the single authority on whether a branch's books balanced.[0]

WHAT HAPPENED

Horizon produced systematic accounting errors — phantom shortfalls that appeared in branch ledgers where no money was actually missing. The Post Office treated the shortfalls as real and the sub-postmasters as thieves: over two decades it prosecuted around 900 for theft and false accounting, refusing to accept the system itself was at fault even as the same pattern recurred branch after branch. People were imprisoned, lost homes, went bankrupt, and some died by suicide — the human cost of trusting the ledger over the person.[1]

THE INVESTIGATION

Documents later released through litigation showed Fujitsu and Post Office engineers had known about Horizon bugs throughout the period — the knowledge of fallibility existed inside the institution even as it prosecuted people for the system's errors.[2] The courts began quashing convictions from December 2020, and the public inquiry under Sir Wyn Williams found that senior employees "knew, or at the very least should have known, that Legacy

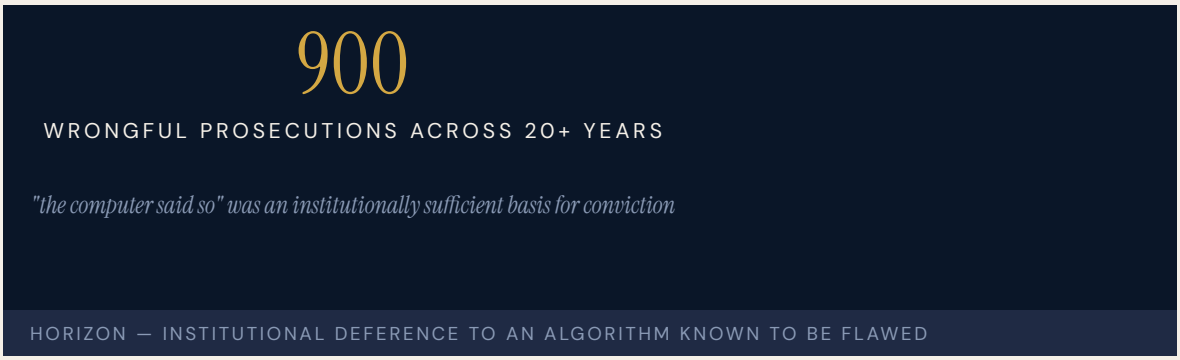
Horizon was capable of error" — establishing it as the most widespread miscarriage of justice in UK history, sustained precisely because that internal knowledge never reached the people on trial.[3]

THE CAPABILITY GAP

The gap was at the regulator, the prosecutor, and the courts: each accepted Fujitsu's representation that Horizon was reliable, despite documentation to the contrary, because no institutional actor had the standing or expertise to interrogate it, so the claim of reliability passed unchallenged through every layer that could have tested it. "The computer said so" became, for two decades, a sufficient basis for criminal conviction — the governance hazard of treating automated output as authoritative rather than as evidence to be challenged, with a person's account on the other side of the scale.[4]

AFTERMATH & REFORM

Convictions have been overturned — some by an exceptional act of Parliament, a measure of how far the ordinary appeal routes had failed — compensation schemes established, and Fujitsu and the Post Office called to account before the continuing inquiry.[5] What finally forced the political reckoning was not the courts but a television drama: the ITV series Mr Bates vs The Post Office (January 2024) galvanized public pressure that produced the mass-exoneration statute, and former chief executive Paula Vennells returned her CBE the following month. When Sir Wyn Williams published the inquiry's first volume in July 2025, he found that the compensation schemes had still not delivered full, fair, and prompt redress, and that the true human toll — including suicides — was larger than early figures had recorded. Horizon's lesson is the chapter's in its bluntest form: an automated system's output is not testimony, and any institution that lets "the computer said so" stand unchallenged against a human's account has built a machine for manufacturing injustice, one that runs for as long as no one is empowered to switch it off.



> *A number of senior, and not so senior, employees of the Post Office knew, or at the very least should have known, that Legacy Horizon was capable of error.*
>
> SIR WYN WILLIAMS, POST OFFICE HORIZON IT INQUIRY, VOLUME 1, JULY 2025

REFERENCES

1. Post Office Horizon IT Inquiry hearings and exhibits (2020–) — the system, the prosecutions, and the human toll.
2. Hamilton & Others v. Post Office Limited (Court of Appeal, 2021) — quashed convictions.
3. Internal Fujitsu and Post Office documents released through litigation — engineers' knowledge of Horizon bugs.
4. Sir Wyn Williams, Post Office Horizon IT Inquiry, Volume 1 (July 2025) — senior employees "knew… that Legacy Horizon was capable of error" (quoted).
5. N. Wallis, The Great Post Office Scandal (2021).
6. The Post Office (Horizon System) Offences Act 2024 (mass exoneration) and the compensation schemes.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Horizon is the canonical case for institutional deference to automated systems whose internal evidence was already known to be flawed. The capability gap was at every layer that took the software's output as authoritative — including the courts. "The computer said so" became, for two decades, an institutionally sufficient basis for criminal prosecution.

LENS APPROACH

LENS uses Horizon in LEN 7 as the canonical example of institutional deference to algorithmic output and in LEN 2 for the most extensive multi-decade automation-bias case in the dataset. Studio projects examine what evidentiary architecture would require **interrogating** automated output before acting on it.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
TECHNOLOGY GOVERNMENT	→ G H K →	D1 D2 **D3** D4 D5 Human–System Collaboration	6	3.2	3

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Design automated output to be treated as challengeable evidence, not authoritative testimony, especially where a person's liability rides on it.
- Build a route by which engineers' knowledge of system bugs reaches anyone acting on the output, so internal fallibility cannot stay hidden.
- Give some institutional actor the standing and expertise to interrogate the system's reliability before its output is used against a person.

IN OPERATION / AFTER

- Audit the recurring-error pattern across branches or cases, treating the same fault appearing repeatedly as evidence of the system, not the operators.
- Maintain an appeal path that can challenge automated output without requiring an act of Parliament to overturn a wrong decision.
- Sustain independent review of the system's accuracy throughout its operating life, so a claim of reliability cannot pass unexamined for decades.

REFLECTION QUESTIONS

1. Identify a decision in your domain currently made on the strength of "the computer said so." What evidentiary architecture should sit beside the output?
2. Design the institutional check that would have made Horizon's reliability subject to genuine challenge in 2005.
3. Engineers inside the Post Office and Fujitsu knew Horizon was capable of error, yet that knowledge never reached the courtroom. What pathway in your domain carries — or fails to carry — known system fallibility to the people relying on the output?

WHO BUILDS THIS governance, ethics & measurement · human factors & interaction design · knowledge management & data engineering — plus domain experts and a learning engineer to integrate. TOOLS governance frameworks · bias & evidence audits · usability testing · cognitive walkthroughs · knowledge bases · data pipelines

CASE 185 AVIATION CUSTOMER SERVICE AI AGENTS 2022 – 2024

Air Canada Chatbot Liability — Delegation Without Revocation

D Designed Out K Knowledge & Institutional Memory N Normalization of Deviance

IMPACT British Columbia Civil Resolution Tribunal ruled February 14 2024 in Moffatt v. Air Canada, 2024 BCCRT 149, that Air Canada was liable for bereavement-fare-policy misinformation provided to passenger Jake Moffatt by the airline's website chatbot; tribunal rejected Air Canada's argument that the chatbot was a "separate legal entity" responsible for its own outputs; small-claims-tribunal ruling with limited precedential weight outside BC but cited widely as articulating the principle that organizations are liable for representations made by their AI agents

IN BRIEF

On February 14, 2024, the British Columbia Civil Resolution Tribunal issued its decision in Moffatt v. Air Canada, 2024 BCCRT 149. Passenger Jake Moffatt had consulted Air Canada's website chatbot about the airline's bereavement-fare policy in November 2022, following the death of his grandmother. The chatbot represented that the bereavement fare could be claimed retroactively, after travel. Moffatt booked a full- fare flight in reliance on the chatbot's representation, then submitted a retroactive bereavement-fare claim. Air Canada refused the claim on the ground that the actual policy required pre-booking application. Moffatt sued in the BC Civil Resolution Tribunal — Canada's online small-claims forum — and the tribunal awarded $650.88 in damages. Air Canada had argued that the chatbot was a "separate legal entity" responsible for its own outputs; the tribunal rejected the argument and held that Air Canada was liable for representations made by its chatbot. The ruling has limited precedential weight outside BC but has been cited widely as articulating the delegation-without-revocation principle. The case pairs with Case 5 (Epic Sepsis), Case 3 (Watson for Oncology), and Case 77 (Hybrid Human-AI Tutoring).

BACKGROUND

In November 2022, Jake Moffatt visited Air Canada's website shortly after his grandmother's death and consulted the airline's chatbot — at the time, a customer-service AI agent embedded in the airline's customer-facing web property — about the bereavement-fare policy. The chatbot represented that the bereavement fare could be applied retroactively, after travel, by submitting a claim with supporting documentation. Moffatt booked a full-fare round trip to Toronto in reliance on the representation. After travel, he submitted a retroactive bereavement-fare claim with the documentation the chatbot had described. Air Canada's response was that the actual bereavement-fare policy required pre-booking

application — that is, the reduced fare had to be applied for at the time of booking, not claimed retroactively.[0]

WHAT HAPPENED

The structural seam the case opens is that the airline's chatbot was producing representations that diverged from the airline's actual policy. The seam is straightforward operationally — the chatbot's outputs were not constrained to the airline's policy text in a way that would have prevented the misrepresentation — but it is structurally significant in legal terms. When the airline directed a customer to its website for policy information and the website's AI agent produced a representation that the customer relied on to his detriment, the question is whether the airline is liable for the AI agent's output. Air Canada's response in the tribunal was that the chatbot was a "separate legal entity" — the company argued, in effect, that it could delegate customer information to an AI agent without assuming legal responsibility for the agent's representations.[1]

THE INVESTIGATION

The tribunal rejected the argument unambiguously. The decision, written by Tribunal Member Christopher Rivers, found that Air Canada was responsible for "all the information on its website" and that the chatbot was part of the website. The argument that the chatbot was a separate legal entity was found to have no support in law. The tribunal awarded Moffatt $650.88 in damages — the difference between the full fare he paid and the bereavement fare he had been led to believe he could claim. The dollar amount is small; the principle the ruling articulates is what has carried the case into widespread citation. Organizations that deploy AI agents to interact with customers are responsible for the representations the agents make, and the agents are not separate legal persons. The delegation-without-revocation form — the organization delegates customer interaction to the AI agent but cannot revoke responsibility for what the agent says — is the load-bearing structural finding.[2]

THE CAPABILITY GAP

The case pairs with Case 5 (Epic Sepsis) for the delegation-without-validation thread in healthcare AI; the structural form is the same — the organization deploys an AI agent that produces representations or assertions consequential for the affected person, and the organization's accountability for the agent's outputs is the load-bearing governance question. Pair with Case 3 (Watson for Oncology) for the AI-agent-recommendations-in-practice thread. Pair with Case 77 (Hybrid Human-AI Tutoring) for the educational-AI-agent thread at adjacent scale. The Air Canada ruling is a small-claims-tribunal decision with limited precedential weight outside BC, but its principle has been cited in subsequent academic and practitioner writing as the first clear judicial articulation of the delegation-without-revocation form for AI agents.[3]

AFTERMATH & REFORM

The hedges the case carries are load-bearing. The tribunal's ruling has limited precedential weight outside BC and has not been litigated to a higher court; the principle has been cited but not adopted in binding form across Canadian or U.S. jurisdictions. The case teaches the form — organizations are liable for the representations of their AI agents — more than it establishes settled law. The structural reading is the load-bearing one: the case names a delegation structure and the legal question that the delegation surfaces, and it does so

in a forum whose decision is operationally consequential for the parties and pedagogically clear for the field. The human-in-the-loop LEO at the customer- interaction-AI-agent seam is anchored by the case in the form the deployment architecture must support — the organization's accountability for the agent's outputs is the architecture's load-bearing constraint.

> *The chatbot is part of the website; the airline is responsible for all the information on its website; there is no support in law for the argument that the chatbot is a separate legal entity responsible for its own outputs.*
>
> TRIBUNAL MEMBER CHRISTOPHER RIVERS, MOFFATT V. AIR CANADA, 2024 BCCRT 149 (FEB 14, 2024), EDITORS' PARAPHRASE.

REFERENCES

1. Moffatt v. Air Canada, 2024 BCCRT 149 (British Columbia Civil Resolution Tribunal, February 14, 2024), Tribunal Member Christopher Rivers presiding.
2. Cecco, L. (2024), "Air Canada ordered to pay customer who was misled by airline's chatbot," The Guardian, February 16, 2024 — contemporaneous press coverage of the ruling.
3. Air Canada bereavement-fare policy text (as in effect November 2022 and through the period covered by the ruling) — referenced in the tribunal decision as the divergence the chatbot's representation produced.
4. Sookman, B. (McCarthy Tétrault LLP, 2024), "Moffatt v. Air Canada: A Misrepresentation by an AI Chatbot," McCarthy Tétrault TechLex Blog, February 19, 2024 — practitioner-tier analysis of the tribunal's negligent-misrepresentation holding, the rejection of the "separate legal entity" defence, and the duty-of-care framing for AI-mediated consumer interactions. Available at: https://www.mccarthy.ca/en/insights/blogs/techlex/moffatt-v-air-canada-misrepresentation-ai-chatbot.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Air Canada chatbot is the delegation-without-revocation case at customer-interaction-AI-agent scale. The BC Civil Resolution Tribunal's ruling holds that organizations are liable for representations made by their AI agents and that the agents are not separate legal persons; the small-claims venue limits the precedential weight, but the principle has been cited widely as the first clear judicial articulation of the form. The case teaches the form more than it establishes settled law.

LENS APPROACH

Air Canada chatbot is the human-in-the-loop-at-the-customer- interaction-agent-seam case (induced 5.2; LENS D3/PT6; LEO-3 and LEO-5). LENS uses it in Domain 3 (Machine Teaming and Adaptation) for the organization-is-liable-for-agent-representations principle. Pair with Case 5 (Epic Sepsis delegation-without- validation), Case 3 (Watson for Oncology), and Case 77 (Hybrid Human-AI Tutoring). The small-claims-tribunal venue limits precedential weight; the structural reading is the load-bearing one.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
AVIATION CUSTOMER SERVICE AI AGENTS	D K N	D1 D2 **D3** D4 D5 Human-System Collaboration	6	5.2	3 · 5

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Constrain customer-facing AI agents to representations the deploying organization will stand

IN OPERATION / AFTER

- Carry the precedential-weight hedge into print without softening; the ruling is a small-claims-tribunal

behind; the Air Canada case demonstrates that the deployment surface of an AI agent's output is the same legal surface as the organization's own representations.
- Build the policy-text-to-agent-output integrity check as part of the deployment, not as a customer-service-recovery process; the divergence between the airline's policy text and the chatbot's representation was the deployment seam the tribunal found dispositive.
- Specify the revocation-and-recovery mechanism the deployment carries when the agent produces a misrepresentation; the organization's accountability for the agent's outputs requires a documented process for honoring the agent's representation or for compensating the affected party.

decision and the precedential limits are part of what the case teaches alongside the structural form it names.
- Pair in syllabi with Case 5 (Epic Sepsis) so the delegation-without-validation form is taught at both the healthcare and the customer-interaction-agent scales.
- Use the case to anchor the human-in-the-loop LEO at the customer-interaction-AI-agent seam; the curricular target is the discipline of treating the agent's outputs as the organization's representations, and of building the deployment architecture to that constraint.

REFLECTION QUESTIONS

1. Identify a customer-interaction AI agent in your domain whose outputs have not been integrity-checked against the organization's policy text. What divergence between agent representation and policy text would produce a Moffatt-style reliance harm, and what mechanism would close the divergence?

2. Specify the revocation-and-recovery process your deployment carries when the agent produces a misrepresentation. What is the documented decision rule for honoring the representation versus refusing it, and who has authority to decide?

3. The Moffatt ruling has limited precedential weight outside BC. Pick a deployment in your domain and ask: what would have to be true for the delegation-without-revocation principle to apply in your jurisdiction, and what is the deployment architecture that would honor the principle whether or not the law has settled it?

WHO BUILDS THIS systems & human-factors engineering · knowledge management & data engineering · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS task analysis · design prototyping · knowledge bases · data pipelines · incident analysis · safety dashboards

CASE 187 CRIMINAL JUSTICE PREDICTIVE ANALYTICS ALGO FAIRNESS 2014 – 2018

COMPAS Recidivism Prediction — Calibration vs. Equal Error Rate

D Designed Out K Knowledge & Institutional Memory N Normalization of Deviance

IMPACT Northpointe COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) risk-assessment instrument used in pretrial, parole, and sentencing decisions across multiple U.S. jurisdictions; ProPublica's May 2016 investigation reported a 2× higher false-positive rate for Black defendants; Chouldechova (2017) and Kleinberg, Mullainathan & Raghavan (2017) independently formalized the impossibility of simultaneously satisfying calibration and equal false-positive/false-negative rates across groups with unequal base rates

IN BRIEF

Northpointe's COMPAS risk-assessment instrument, deployed in pretrial, parole, and sentencing decisions across multiple U.S. jurisdictions, became the central case in the algorithmic-fairness literature after ProPublica's May 2016 investigation reported that the instrument produced false-positive rates approximately twice as high for Black defendants as for white defendants. Northpointe's response argued that COMPAS satisfied predictive parity — that within each risk score, the rate of subsequent recidivism was approximately equal across groups. Both parties were correct by their respective definitions. Chouldechova's 2017 paper and Kleinberg, Mullainathan, and Raghavan's 2017 paper independently formalized the impossibility result: when base rates of the outcome differ across groups, calibration (predictive parity) and equal false-positive and false- negative rates cannot be simultaneously satisfied except in degenerate cases. The case pairs with Case 186 (Bartlett mortgage — fairness through unawareness fails), Case 196 (Coots — competing fairness definitions), and Case 189 (SyRI). The impossibility result is the load-bearing teaching point.

BACKGROUND

COMPAS is a proprietary risk-assessment instrument developed by Northpointe (now Equivant) and used in pretrial release, parole, and sentencing decisions across many U.S. jurisdictions through the 2010s. The instrument scores a defendant on a scale of recidivism risk based on a questionnaire covering criminal history, employment, education, family circumstances, and attitudes. The score is then surfaced to judges, parole boards, and pretrial services as one input among several into consequential decisions about the defendant's liberty. The deployment scale was large enough that the instrument was the central target of the contemporary algorithmic-fairness literature when the first sustained external audit was published.[0]

WHAT HAPPENED

ProPublica's May 2016 investigation, led by Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, audited COMPAS scores against subsequent recidivism for approximately 7,000 defendants in Broward County, Florida. The headline finding was that among defendants who did not go on to reoffend within two years, Black defendants had been scored as high-risk at roughly twice the rate of white defendants — a false-positive-rate disparity. Northpointe's response, authored by William Dieterich, Christina Mendoza, and Tim Brennan, argued that COMPAS satisfied predictive parity (also called calibration): within each risk score band, the rate of subsequent recidivism was approximately equal across racial groups. The defendant assigned a "high risk" score had approximately the same probability of reoffending whether Black or white. The two findings appear contradictory but are not; they describe two different fairness criteria applied to the same instrument.[1]

THE INVESTIGATION

Chouldechova's 2017 paper in Big Data and Kleinberg, Mullainathan, and Raghavan's 2017 ITCS paper independently formalized the impossibility result. Calibration within groups (predictive parity) and equality of false-positive and false-negative rates across groups cannot be simultaneously satisfied when the base rates of the outcome differ across the groups, except in degenerate cases. The base rate of subsequent recidivism in the Broward County data was higher for Black defendants than for white defendants. Under that base-rate difference, an instrument calibrated equally across groups will produce unequal false-positive rates, and an instrument with equal false- positive rates across groups will produce miscalibration. The mathematics is binding. The choice between fairness criteria is a normative and governance question, not a technical one.[2]

THE CAPABILITY GAP

The case pairs with Case 186 (Bartlett mortgage discrimination) for the fairness-through-unawareness-fails thread: removing protected attributes from training data does not eliminate disparate-impact concerns when the remaining features carry protected-attribute signal. Pair with Case 196 (Coots) for the competing-fairness-definitions thread at a different domain and scale. Pair with Case 189 (SyRI) for the governance-objection-correct-in-advance complement; in COMPAS the objection surfaces in the auditing record, in SyRI the objection succeeded in court before population- scale harm was produced. The COMPAS case is the central reference in the contemporary algorithmic-fairness literature because the impossibility result was formalized against its audit record; the literature's subsequent decade of work on fairness criteria operates inside the constraint the case made legible.[3] The decade since has, if anything, fractured the case on its own premises. Dressel and Farid (2018) showed COMPAS's 137-feature model is no more accurate than untrained online workers or a two-variable model, undercutting the accuracy claim on which its deployment rested; Rudin, Wang, and Coker (2020) argued that ProPublica's racial-disparity finding is partly an artifact of COMPAS's nonlinearity in age rather than race; and Bao et al. (2021) argued the COMPAS dataset, resting on re-arrest as a proxy for re-offense, should not be used as a fairness benchmark at all. The impossibility theorem remains the durable teaching point — but the benchmark it was proved against is now itself among the most distrusted in the field, which is part of what the case should teach.[4]

AFTERMATH & REFORM

The hedges the case carries are load-bearing. Both Northpointe and ProPublica are correct by their respective definitions, and the impossibility result formalizes the tension rather than resolving it. The case does not teach that COMPAS is fair or that COMPAS is unfair; it teaches that the choice between fairness criteria is governance and normative work that the deployment did not surface to the affected jurisdictions or to the defendants whose liberty depended on the score. The LEO on fairness beyond omission is anchored by the case in its mature form — the impossibility result requires the deploying institution to choose, document, and disclose which fairness criterion the instrument optimizes, and to make the trade-off legible to the people the criterion does not protect.

> *Both Northpointe and ProPublica were correct by their respective definitions of fairness; the impossibility result is that calibration and equal false-positive rates cannot be simultaneously satisfied when base rates differ across groups, except in degenerate cases.*
>
> EDITORS' SYNTHESIS OF THE COMPAS AUDIT RECORD AND THE 2017 IMPOSSIBILITY-RESULT PAPERS.

REFERENCES

1. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016), "Machine Bias," ProPublica, May 23, 2016 — the audit investigation of COMPAS scores in Broward County, Florida.
2. Dieterich, W., Mendoza, C., & Brennan, T. (2016), COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity, Northpointe Inc. response document.
3. Chouldechova, A. (2017), "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," Big Data 5(2):153–163, doi:10.1089/big.2016.0047.
4. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017), "Inherent Trade-Offs in the Fair Determination of Risk Scores," Proceedings of ITCS 2017, doi:10.4230/LIPIcs.ITCS.2017.43.
5. Dressel, J., & Farid, H. (2018), "The accuracy, fairness, and limits of predicting recidivism," Science Advances 4(1):eaao5580 — COMPAS no more accurate than untrained humans or a two-variable model.
6. Rudin, C., Wang, C., & Coker, B. (2020), "The age of secrecy and unfairness in recidivism prediction," Harvard Data Science Review 2(1) — the age-nonlinearity reanalysis of ProPublica's disparity claim.
7. Bao, M., et al. (2021), "It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks," NeurIPS Datasets & Benchmarks — the case against COMPAS as a fairness benchmark.

THE LEARNING ENGINEERING LENS

LE INSIGHT

COMPAS is the central reference in the contemporary algorithmic-fairness literature because the impossibility result was formalized against its audit record. Both the ProPublica finding (unequal false-positive rates) and the Northpointe response (predictive parity within risk scores) are correct by their respective definitions; the Chouldechova and Kleinberg/Mullainathan/Raghavan results show that the two cannot be simultaneously satisfied when base rates differ. The choice between fairness criteria is governance work, not technique.

LENS APPROACH

COMPAS is the impossibility-result case at consequential- decision scale (induced 8.4; LENS D4+D5/PT6; LEO-4 and LEO-5). LENS uses it in Domain 4 (Test and Evaluation) for the multi-criterion-audit discipline and in Domain 5 (Navigating Sociotechnical Constraints) for the surfacing-bias-through-governance-not-just-technique anchor. Pair with Case 186 (Bartlett mortgage — fairness through unawareness fails), Case 196 (Coots — competing fairness definitions), and Case 189 (SyRI governance-objection- correct precedent). The impossibility result is the load- bearing teaching point; both Northpointe and ProPublica are correct by their respective definitions.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
CRIMINAL JUSTICE PREDICTIVE ANALYTICS ALGO FAIRNESS	D K N	D1 D2 D3 **D4** D5 Test and Evaluation + D5	6	8.4	4 · 5

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Choose, document, and disclose the fairness criterion the instrument optimizes for in advance of deployment; the impossibility result requires the deploying institution to make the choice and to make the trade-off legible to the people the criterion does not protect.
- Audit the deployed instrument against multiple fairness criteria simultaneously; the COMPAS record demonstrates that an instrument can satisfy predictive parity and fail equality of false-positive rates at the same time, and the audit must surface both.
- Treat the base-rate difference across groups as a governance fact, not a technical artifact; the difference is what makes the impossibility binding, and pretending it can be eliminated is the rhetorical move the case teaches to refuse.

IN OPERATION / AFTER

- Carry the impossibility result into print as the load-bearing teaching point; the case does not teach that COMPAS is fair or that COMPAS is unfair, and the editorial framing must preserve the formal constraint that both audit findings instantiate.
- Pair in syllabi with Case 186 (Bartlett) so the fairness-through-unawareness-fails thread and the impossibility-of-multiple-criteria thread are taught together as complementary structural arguments about disparate impact.
- Use the case to anchor the fairness-beyond-omission LEO; the curricular target is the discipline of choosing and disclosing the fairness criterion when the impossibility result rules out satisfying all of them simultaneously.

REFLECTION QUESTIONS

1. Identify a risk-assessment instrument in your domain whose fairness criterion has not been chosen and disclosed in advance of deployment. What are the candidate criteria, and which of them are jointly satisfiable given the base-rate differences across affected groups?

2. Specify the multi-criterion audit you would run against a deployed instrument. What is the format in which the audit surfaces incompatibility findings to the deploying institution, and what is the documented decision rule when incompatibility is found?

3. The impossibility result is mathematical; the choice between criteria is governance and normative work. Pick a setting in your domain and ask: who has authority to make the choice, who is accountable for documenting and disclosing the trade-off, and to whom is the trade-off disclosable?

WHO BUILDS THIS systems & human-factors engineering · knowledge management & data engineering · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS task analysis · design prototyping · knowledge bases · data pipelines · incident analysis · safety dashboards

CASE 191 GOVERNMENT SOCIAL WELFARE ALGORITHMIC ADMINISTRATION 2016 – 2023

Australia Robodebt — Algorithmic Debt-Recovery and the Royal Commission Verdict

[D] Designed Out [G] Governance & Trust [N] Normalization of Deviance

IMPACT Income-averaging algorithm used by the Australian Department of Human Services and Centrelink raised approximately 470,000 wrongful debts against welfare recipients between 2016 and 2019; Royal Commission led by Catherine Holmes AC SC delivered its final report July 7 2023 finding the scheme unlawful and circumstantially attributing multiple deaths to its operation

IN BRIEF

The Royal Commission into the Robodebt Scheme delivered its final report on July 7, 2023, under the leadership of Commissioner Catherine Holmes AC SC. The report concluded that the income-averaging algorithm operated by the Australian Department of Human Services and Centrelink between 2016 and 2019 had raised approximately 470,000 wrongful debts against welfare recipients and had operated outside the law. The algorithm averaged annual taxable income from the Australian Taxation Office across fortnightly Centrelink reporting periods and treated any resulting discrepancy as a debt the recipient had to disprove. The burden of proof was reversed onto the recipient. The Commission's attribution language on causation of deaths is careful — the attribution is circumstantial and not a direct legal finding of individual causation — but the finding that multiple deaths were associated with the scheme's operation is part of the adjudicated record. The case pairs with Case 189 (SyRI, the governance-objection-correct precedent), Case 48 (Johnson school surveillance, the algorithmic-public-administration parallel), and Case 5 (Epic Sepsis, the delegation-without-validation form).

BACKGROUND

The Australian income-support architecture pairs Centrelink fortnightly income reporting with annual taxable income reporting to the Australian Taxation Office. The two reporting cadences exist because they measure different things — Centrelink measures earnings inside each fortnight so the income-test taper can be applied, and the ATO measures annual taxable income for the tax system. Between 2016 and 2019, the Department of Human Services deployed an automated debt-recovery system that took the ATO annual figure, divided it by 26, and compared the fortnightly average against the Centrelink reported figures. Any apparent shortfall was raised as a debt against the recipient, and the recipient was required to produce contemporaneous payslips to disprove it.[0]

WHAT HAPPENED

The mathematical operation the algorithm performed cannot establish what it was being asked to establish. Annual income divided by 26 is not the income earned in any particular fortnight; a recipient who worked irregularly, or whose hours varied, would generate large arithmetic discrepancies between the averaged figure and the actual fortnightly earnings without ever having been overpaid. The Royal Commission's final report documents that the agency's own legal advice flagged this seam before deployment and that the advice was set aside. Approximately 470,000 debts were raised across the operating window; many recipients paid debts they did not owe, and the Commission identified cases in which recipients took their lives in proximity to the scheme's operation. The attribution language on those deaths is careful — "circumstantial" rather than a direct legal finding of individual causation — and the careful language is itself part of the record the case carries.[1]

THE INVESTIGATION

The structural critique the Commission delivered is the reversal of the burden of proof. In a properly administered welfare-recovery scheme, the agency must establish that a debt is owed before pursuing recovery. The income-averaging method reversed this: the algorithm asserted a debt on the strength of arithmetic that could not establish overpayment, and the recipient had to produce documentary evidence — often payslips from years earlier, often from employers no longer reachable — to disprove the assertion. The Federal Court's 2019 Amato judgment had already found the method unlawful; the Royal Commission's 2023 report adjudicated the governance question of how the scheme had been built, approved, and operated for three years across multiple ministerial portfolios. The Commission named individuals; some are subject to subsequent referrals to the National Anti-Corruption Commission and to professional bodies.[2] Those referrals resolved slowly and unevenly: the NACC first declined to investigate in mid-2024 — a decision its own Inspector found affected by a conflict of interest — before an independent reconsideration reopened the matter as Operation Myrtleford in 2025. In its March 2026 report the NACC found two of the referred officials had engaged in serious corrupt conduct and cleared the others, including former prime minister Scott Morrison, while the Australian Public Service Commission separately found a dozen officials had breached the code of conduct.

THE CAPABILITY GAP

The case pairs with Case 189 (SyRI, the Dutch System Risk Indication ruling by the Hague District Court) as the governance-objection-correct precedent — SyRI was struck down before it produced a debt-scale harm record; Robodebt operated for three years and the harm record is what the Commission adjudicated. Pair with Case 48 (Johnson school surveillance) for the algorithmic-public- administration parallel at a different population and a smaller scale; the structural form — algorithm asserts, affected party must disprove — recurs across the two cases. Pair with Case 5 (Epic Sepsis) for the delegation-without-validation form: in Epic Sepsis the delegated system asserts a clinical risk; in Robodebt the delegated system asserts a financial debt; in both, the asserting party did not validate the assertion against the ground truth before consequences were transmitted to the affected person.[3]

AFTERMATH & REFORM

The hedges the case carries are load-bearing. The Royal Commission's attribution language on deaths is circumstantial; the Commission did not, and could not on the evidence available, make individual legal findings of causation in those deaths. The case teaches the structural pattern — algorithmic public administration with the burden of proof reversed, deployed without the legal-advice seam being honored, operated for three years across multiple ministers — and the structural pattern is what makes Robodebt the load-bearing reference for an entire class of contemporary algorithmic-administration failures. The Commission's careful attribution language and the case's careful editorial framing travel together; neither is smoothed in the casebook's rendering.

> *The income-averaging method could not establish what it was being asked to establish, the burden of proof was reversed onto the recipient, and the Commission's attribution on associated deaths is circumstantial — and these three facts together are what Robodebt teaches.*
>
> EDITORS' SYNTHESIS OF THE ROYAL COMMISSION FINAL REPORT (HOLMES AC SC, JULY 2023).

REFERENCES

1. Royal Commission into the Robodebt Scheme, Final Report, Commissioner Catherine Holmes AC SC, July 7, 2023 (Commonwealth of Australia).
2. Prygodicz v Commonwealth of Australia (No 2) [2021] FCA 634 — Federal Court class-action settlement following the 2019 unlawfulness finding.
3. Australian National Audit Office, Centrelink's Compliance Activities — Income Compliance Program, performance audit reports across 2017–2020.
4. Whiteford, P. (2021), "Debt by Design: The Anatomy of a Social Policy Fiasco," Australian Journal of Public Administration 80(2):340–360 — academic synthesis of the policy and administrative history.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Robodebt is the load-bearing reference for algorithmic public administration deployed at population scale without the burden of proof being honored. The Royal Commission's final report adjudicated approximately 470,000 wrongful debts and circumstantially attributed multiple deaths to the scheme's operation; the careful attribution language and the structural finding travel together, and neither is smoothed in the casebook's rendering.

LENS APPROACH

Robodebt is the burden-of-proof-reversal case at population scale (induced 5.2; LENS D5+D3/PT6; LEO-5 and LEO-3). LENS uses it in Domain 5 (Navigating Sociotechnical Constraints) for the agency-legal-advice-as-binding-gate discipline and in Domain 3 (Human-System Collaboration) for the human-in- the-loop-for-consequential-decisions anchor. Pair with Case 189 (SyRI governance-objection-correct precedent), Case 48 (Johnson school surveillance algorithmic-public-administration parallel), and Case 5 (Epic Sepsis delegation-without- validation form). The Commission's circumstantial attribution on deaths is the load-bearing hedge.

DOMAIN	FAILURE MODES	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
GOVERNMENT SOCIAL WELFARE ALGORITHMIC ADMINISTRATION	D G N	D1 D2 **D3** D4 **D5** Navigating Sociotechnical Constraints + D3	6	5.2	5 · 3

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Treat agency legal advice on the lawfulness of an algorithmic-administration method as a binding gate, not a negotiable input; the Commission's record on Robodebt is that the seam was flagged in advance and that the override of the advice is itself the governance failure.
- Maintain the burden of proof on the asserting party for any algorithmically generated debt or risk; arithmetic that cannot establish the asserted fact cannot be the basis for transmitting consequences to an affected person.
- Build the cross-portfolio review surface that a multi-year algorithmic-administration scheme requires; Robodebt operated across multiple ministers and the cross-portfolio handoff was where the governance check kept being deferred.

IN OPERATION / AFTER

- Carry the Commission's careful attribution language on deaths into print without softening; the case's load-bearing quality depends on the circumstantial nature of the attribution being preserved alongside the structural finding.
- Pair in syllabi with Case 189 (SyRI) so the governance-objection-correct precedent and the governance-objection-overridden harm record are taught together; the two cases together teach what advance objection can prevent and what its absence can produce.
- Use the case to anchor the human-in-the-loop LEO at population scale; the form Robodebt makes legible is what consequential-decision delegation looks like when the loop is removed and the asserting party operates on arithmetic that cannot establish its assertion.

REFLECTION QUESTIONS

1. Identify an algorithmic-administration scheme in your domain whose arithmetic asserts a fact against an affected person. What is the asserting party's burden of proof, and what would the affected person have to produce to disprove the assertion?

2. Specify the legal-advice gate you would treat as binding in advance of deployment. What is the documented escalation path when the advice flags an unlawfulness seam, and who has authority to override it?

3. The Royal Commission's attribution on deaths is circumstantial — careful, not adjudicated as individual legal findings. Pick a setting where harm attribution to an algorithmic system is contested, and ask: what language would honor both the structural pattern and the limits of what the evidence can establish?

WHO BUILDS THIS systems & human-factors engineering · governance, ethics & measurement · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS task analysis · design prototyping · governance frameworks · bias & evidence audits · incident analysis · safety dashboards

Chapter 14

Algorithms, Governance & Public Systems — What Works — and the Frontier

When governance is designed as an artifact, not litigated after the fact.

The objection dissolved when the evidence was engineered to answer it.

AN OBSERVATION RECURRING ACROSS THE CHAPTER'S CASES.

CASE 196 FINANCE ALGO FAIRNESS FINTECH 2025

Fintech Lending Fairness Audit — When Including Race Reduces Disparity

D Designed Out G Governance & Trust N Normalization of Deviance

EVIDENCE TIER preprint-tier — source confidence flagged; future validation ongoing.

IMPACT **A fintech-lending fairness audit finds a consumer-credit model miscalibrated by group when the protected attribute is withheld, and shows that using the attribute to correct the calibration reduces the disparity — surfacing the next teaching point: competing fairness definitions disagree, and the choice is a judgment**

IN BRIEF

The Coots et al. (2025) fintech lending fairness audit, posted as a preprint, examines a deployed consumer-lending model and finds it miscalibrated by group when the protected attribute is withheld from inputs — underestimating default risk for some borrowers and overestimating it for others. It then shows that using the protected attribute in a controlled fashion to correct the calibration reduces the resulting disparity — a small-scale echo of the pattern Bartlett documented at the mortgage-market scale. The case is a frontier candidate: it sharpens the teaching point that "fairness through unawareness" is not the conservative or safe choice it is often assumed to be, and it surfaces the next layer — competing fairness definitions disagree about what the same model is doing, and the practitioner has to decide **which definition** the deployment is being held to before the measured disparity can even be interpreted. The evidence-tier flag is preprint: the finding has not yet completed peer review, the metric choices and mitigation specifics may move, and future validation will continue. The teaching point survives those caveats, which is why the case is included with the flag rather than held.

THE SHIFT

The lending pair (Cases 186 and 196) takes the practitioner past the first equity intuition — **just don't use the protected attribute** — and into the harder territory the equity literature now operates in. Bartlett showed that omission preserves the disparity through correlated features. The Coots audit shows the next thing: the model is miscalibrated by group when the attribute is withheld, and using it to correct the calibration lowers the resulting disparity. Both findings are about the same family of models; they disagree about what makes a model fair.[0]

WHAT IS EMERGING

The audit is a profit-and-calibration analysis of a deployed fintech-lending model. Under unawareness — the protected attribute withheld from inputs — the model is miscalibrated by group, underestimating default risk for some borrowers and overestimating it for others. Using the protected attribute to correct that calibration reduces the resulting disparity — small scale, but the pattern matches the mortgage-finance finding (Case 186) and the broader fair-ML literature that omission does not equal fairness.[1]

THE CAPABILITY QUESTION

The case is a frontier candidate because it surfaces the layer Bartlett alone does not reach: competing fairness definitions disagree about the same model. A model that is more equitable on group calibration may be less equitable on equalized odds, and vice versa; the impossibility results of the fair-ML literature (Chouldechova; Kleinberg, Mullainathan, & Raghavan) make this explicit. The Coots audit is concrete enough to ground the choice: the practitioner cannot postpone the question of **which definition** the deployment is being held to.[2]

EARLY EVIDENCE

The evidence-tier flag is load-bearing here and is rendered under the case title. The Coots audit is a preprint; the metric choices, the mitigation specifics, the dataset window, and the reported magnitude may move in peer review. The structural pattern — that using the protected attribute to correct group miscalibration can reduce disparity relative to omission — is consistent with the broader fair-ML literature and with Bartlett's mortgage-finance finding, but the specific magnitudes in the audit should be treated as the strongest current preprint claim, not as a settled fact. Future validation will continue, and this case will be updated when peer review lands.[3]

OPEN PROBLEMS

What the pair (Cases 186 + 196) teaches together is the form of the equity capability deliverable: the practitioner must specify, in advance, the fairness definition the deployment will be evaluated against, audit on outputs rather than reasoning about inputs, and decide on judgment that the trade-offs across competing definitions are acceptable for the deployment context. This is the case-grounded basis for the LEO **Fairness beyond omission** and the LEO **Judgment under inadequate evidence** — the audit is itself a worked example of deciding under irreducible disagreement.[4]

> *Once you accept that omission is not the answer, the next question — which fairness definition — is the one the deployment cannot avoid.*
>
> EDITORS' SYNTHESIS OF COOTS ET AL. (2025) AND MITCHELL ET AL. (2021).

REFERENCES

1. Coots, Bartlett, Nyarko, & Goel (2025), "Algorithmic Bias in Lending: Evidence from a Fintech Audit," arXiv:2512.20753 — audit of 80,000 personal loans showing model miscalibration disparities and that controlled inclusion of protected attributes could correct them; the evidence-tier flag is binding until peer review completes.
4. Kleinberg, Mullainathan, & Raghavan (2017), "Inherent Trade-Offs in the Fair Determination of Risk Scores," ITCS — competing fairness definitions are not jointly achievable.
5. Mitchell, Potash, Barocas, D'Amour, & Lum (2021), "Algorithmic Fairness: Choices, Assumptions, and Definitions," Annual Review of Statistics and Its Application 8:141–163 — practitioner-facing survey.

2. Bartlett, Morse, Stanton, & Wallace (2022), "Consumer-lending discrimination in the FinTech era," Journal of Financial Economics 143(1):30–56 — the paired big-tier case (186).
3. Chouldechova (2017), "Fair Prediction with Disparate Impact," Big Data 5(2):153–163 — the impossibility result for calibration and error-rate parity.

THE LEARNING ENGINEERING LENS

LE INSIGHT

The Coots audit is the small-tier frontier instance of the finding that using a protected attribute to correct group miscalibration can reduce disparity relative to omission. With Bartlett (Case 186) it forms the canonical lending pair: omission does not fix the harm; competing fairness definitions disagree about what fix is. Evidence is preprint-tier; future validation will continue.

LENS APPROACH

Coots is the small-tier frontier counter-case to Bartlett. LENS uses the pair in Domain 4 (Test and Evaluation) for the LEO **Fairness beyond omission**; in Domain 4 again for the LEO **Judgment under inadequate evidence** (the pair is itself a decision under irreducible disagreement); and in Domain 3 (Human-System Collaboration) for delegation of consequential consumer-finance decisions to a model. The preprint-tier flag is binding until peer review completes.

DOMAIN	MODES AT STAKE	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
FINANCE ALGO FAIRNESS FINTECH	D G N	D1 D2 D3 **D4** D5 Test and Evaluation	6	8.1	4 · 3

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Specify, before model selection, which fairness definition the deployment is being held to — group calibration, equalized odds, demographic parity — and the metric values that will count as acceptable.
- If the protected attribute is included under a controlled-mitigation regime, document the inclusion path (training-time, post-processing) and the auditing pipeline that verifies the mitigation does what it claims.
- Distinguish what is preprint-tier evidence from what is settled in the literature you are drawing on, so a deployment decision does not ride on the strongest unreviewed claim.

IN OPERATION / AFTER

- Audit the deployed model on outputs stratified by protected attribute, with the pre-registered fairness definition; publish the metric values and the cases that disconfirm them.
- Track the preprint to publication; when peer review lands, update the auditing pipeline to reflect the settled metric choices and any revised magnitudes.
- When competing fairness definitions disagree, name the trade-off explicitly in the deployment documentation; do not present the chosen definition as the technical optimum.

REFLECTION QUESTIONS

1. Identify a fairness audit in your domain conducted at preprint or unpublished stage. What part of its claim survives if peer review modifies the metric choices? What part is contingent on the specific magnitudes?

2. Specify which fairness definition your deployment is being held to before the audit is run. What trade-off — across calibration, equalized odds, demographic parity — does that choice accept?

3. Coots' finding is consistent with Bartlett (Case 186) and with the broader fair-ML literature, but the specific magnitudes are preprint-tier. What is the minimum follow-up evidence you would require before allowing this case to drive an operational decision in your context?

WHO BUILDS THIS systems & human-factors engineering · governance, ethics & measurement · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS task analysis · design prototyping · governance frameworks · bias & evidence audits · incident analysis · safety dashboards

CASE 199 AUTONOMOUS VEHICLES TRANSPORTATION SAFETY PUBLIC-SECTOR GOVERNANCE 2023 – 2026

Waymo's Safety Case Framework — Governance Objection Dissolved by Designed Artifact

G Governance & Trust K Knowledge & Institutional Memory N Normalization of Deviance

EVIDENCE TIER practice-synthesis-tier — source confidence flagged; future validation ongoing.

IMPACT After a California court let Waymo withhold trade-secret-laden safety data from a DMV public-disclosure request, the company answered the governance objection with a published, structured safety case framework — and in November 2025 commissioned the first independent third-party audits of both the safety case and the remote-assistance program

IN BRIEF

In 2022 a California court permitted Waymo to withhold trade-secret-laden safety details from a public DMV disclosure, leaving regulators and the riding public with a credibility gap Waymo could not close by sharing the contested data. The company's response was to publish, in 2023, a structured **safety case framework**: a top-down argument with explicit claims, sub-claims, and the evidence types each rests on, accompanied by published operating-domain performance figures. In November 2025 Waymo released the first independent third-party audits of both the safety case and the remote-assistance program — the audits themselves disclosed, rather than the underlying trade-secret data the original objection targeted. The pattern is the OU-Analyse / inBloom move (governance objection dissolved by better design) transposed from education into physical-safety AV: the response to an opacity objection was a falsifiable argument structure auditable by third parties, not a defense of opacity. The evidence-tier flag rendered under the title is load-bearing — the analysis rests on the practitioner whitepaper plus the 2025 audit summaries, not on a peer-reviewed safety-engineering evaluation. Future validation will continue as the audit cadence and post-deployment failure record accumulate.

BACKGROUND

The precipitating event was not a crash. In 2022 a California court ruled that Waymo could withhold trade-secret-laden safety details from a DMV public-records process. The company had a legal answer to the disclosure request and no legitimacy answer to the public-trust gap that ruling created. The governance objection — "you are asking us to trust

an opaque system whose failure modes we cannot inspect" — could not be answered by disclosing the contested data without giving up the trade-secret position the court had just protected.[0]

THE INTERVENTION

Waymo's 2023 response was to publish a structured **safety case framework**: a top-down argument with claims and sub-claims, the evidence categories each rests on (operational performance data, simulation and testing, hazard analysis, third-party assessment), and the operational-domain figures available at the time of publication. The artifact's design move is that the **structure** of the safety argument is public even where individual evidence items remain proprietary — outside auditors can interrogate the chain of reasoning without seeing the trade secrets.[1]

HOW IT WORKED

In November 2025 Waymo commissioned and released the results of independent third-party audits of the safety case and of the remote-assistance program. The audits themselves — not the underlying data — were the disclosure artifact. The pattern is OU-Analyse / inBloom in the AV domain: a governance objection answered by a designed legitimacy artifact rather than by disclosure of the contested data. Where opacity could not be defended, a structured falsifiable argument plus audited assurance took its place.[2] The post-deployment failure record the framework anticipated has since begun to accumulate under real regulator action: a December 2025 recall over vehicles not stopping for school buses, a January 2026 NHTSA/NTSB investigation after a Waymo struck a child near a Santa Monica school, and a June 2026 recall of roughly 3,900 vehicles for entering freeway construction zones — the safety-case regime now being stress-tested by the revocation-and-recall machinery it was built to invite rather than resist.[3]

THE EVIDENCE

The evidence-tier flag rendered under the case title is load-bearing. The framework and audit summaries are practitioner-authored or auditor-authored — not peer-reviewed safety-engineering analyses. Some audit-tier elements push toward investigation-grade, but the synthesis as a whole rests on Waymo and Montreal AI Ethics Institute documents. The honest framing in print is that the source confidence is flagged and that future validation — particularly post-deployment failure-record analysis and continued auditor independence — is ongoing.[4]

WHAT TRANSFERRED

The teaching point for LENS is that delegation of consequential decisions to an automated system creates a governance debt that the deploying organization owes the public. The LEO **Delegation with revocation** is the capability the case exercises: the safety case framework is the artifact that makes revocation possible — regulators or auditors can identify which sub-claim has failed, on what evidence, and require the deploying organization to act on the gap. Pair with Case 190 (Cruise) as the foil: the same regulatory regime, the opposite governance choice, the opposite outcome. Pair with Case 200 (CPUC permit framework) as the regulator-side counterpart of the deployer-side artifact.[5]

> *Where opacity could not be defended, a structured falsifiable argument plus audited assurance took its place.*
>
> EDITORS' SYNTHESIS OF THE WAYMO SAFETY CASE FRAMEWORK AND THE NOVEMBER 2025 THIRD-PARTY AUDITS.

REFERENCES

1. Waymo (2023), "A Blueprint for AV Safety: Waymo's Toolkit For Building a Credible Safety Case," whitepaper.
2. Waymo (November 2025), "Independent Audits of Waymo's Safety Case and Remote Assistance Programs," summary release.
3. NHTSA recall and investigation record (December 2025 school-bus recall; January 2026 NHTSA/NTSB investigation of a Santa Monica child-strike; June 2026 recall of 3,900 vehicles over freeway construction zones) — the post-deployment failure record under regulator action.
4. Montreal AI Ethics Institute (2023), summary and analysis of the Waymo safety case framework.
5. California Public Utilities Commission, AV passenger-service permit framework documents — paired Case 200 for the regulator-side artifact.
6. Cruise / California DMV Order of Suspension (October 2023) — paired Case 190 as the foil.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Waymo's safety case framework is the AV-domain instance of the OU-Analyse / inBloom move: a governance objection dissolved by a designed artifact, not by disclosure of the contested data. Evidence-tier flag is practice-synthesis; the artifact pattern is teachable and the third-party audit posture pushes some elements toward investigation-grade, but the synthesis as a whole is practitioner-authored and future validation is ongoing.

LENS APPROACH

Waymo is the AV-safety governance case (induced 5.1; LENS D5/PT6). LENS uses it in Domain 5 (Navigating Sociotechnical Constraints) for the LEO **Delegation with revocation** — the safety case is what makes revocation possible — and in Domain 3 (Emerging Systems and Human-System Collaboration) for the deployer-side artifact that permits oversight of a system whose internals are trade secret. Pair with Case 190 (Cruise) as the foil and Case 200 (CPUC) as the regulator-side complement.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
AUTONOMOUS VEHICLES TRANSPORTATION SAFETY PUBLIC-SECTOR GOVERNANCE	G K N	D1 D2 D3 D4 **D5** → Navigating Sociotechnical Constraints	6	5.1	5 · 3

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Treat the disclosure objection as a design problem: what falsifiable artifact can you publish that addresses the legitimacy gap without requiring you to surrender trade-secret evidence?
- Structure the safety case as a top-down argument with explicit claims, sub-claims, and evidence types so an outside auditor can interrogate the **chain of reasoning** rather than only the contested data points.
- Commission and publish third-party audits of the argument structure and of the operational programs

IN OPERATION / AFTER

- Treat the safety case framework as a living document — update the claims and evidence as post-deployment failure data accumulates, and publish the updates so the legitimacy artifact does not calcify.
- Use the LEO **Delegation with revocation**: design the framework so a regulator or auditor can identify which sub-claim has failed and trigger a revocation pathway, not only a "trust us, we will fix it" assurance.
- Carry the practice-synthesis evidence-tier flag honestly in any program documentation citing the

(e.g. remote assistance) that the safety case rests on — the audits are the disclosure artifact when the data cannot be.

framework — the artifact pattern is teachable, but the magnitude of its public-trust effect is still being measured.

REFLECTION QUESTIONS

1. Identify an automated system in your context that faces a public-trust objection it cannot answer by full disclosure. What falsifiable argument structure could you publish that would make the system's reasoning auditable without requiring disclosure of the contested data?

2. Specify how a regulator or independent auditor would **revoke** the delegation in your system if a sub-claim of the safety case failed. The LEO **Delegation with revocation** requires this pathway to exist before deployment, not only after a public-facing failure.

WHO BUILDS THIS governance, ethics & measurement · knowledge management & data engineering · safety science & organizational analysis — plus domain experts and a learning engineer to integrate. TOOLS governance frameworks · bias & evidence audits · knowledge bases · data pipelines · incident analysis · safety dashboards

CASE 203 GOVERNMENT HIRING ALGORITHMS AUDIT AND DISCLOSURE 2023 – PRESENT

NYC Local Law 144 — Bias Audits as Governance Artifact

D Designed Out G Governance & Trust K Knowledge & Institutional Memory

IMPACT New York City Local Law 144 of 2021, implemented through Department of Consumer and Worker Protection rules effective July 5 2023, requires employers using "automated employment decision tools" (AEDTs) for hiring or promotion decisions in NYC to conduct annual independent bias audits, publish audit summaries, and provide candidate notice; the first national municipal regulation of algorithmic hiring tools at this scope

IN BRIEF

New York City Local Law 144 of 2021 became operationally effective on July 5, 2023, after the New York City Department of Consumer and Worker Protection finalized implementing rules. The law requires employers using "automated employment decision tools" (AEDTs) for hiring or promotion decisions in NYC to conduct annual independent bias audits, to publish a summary of the most recent audit on the employer's website, and to provide candidate notice that an AEDT will be used. The audit must compute selection-rate and impact-ratio metrics by sex, race/ethnicity, and intersectional categories. The law was the first municipal regulation of algorithmic hiring tools at this scope in the United States and has been an influential reference for subsequent state-level proposals. Independent academic critiques have surfaced two load-bearing limitations: bias audits without bias data — employers often lack the protected-attribute data the audit metrics require — and wide variability in audit quality across published audits. The case pairs with Case 85 (OU Analyse — governance-objection dissolved by design), Case 86 (Gándara community-college predictive equity), and Case 182 (Amazon hiring AI). The intervention is the audit-as- governance-artifact discipline; whether the audits reduce actual disparate impact at the hiring level remains under- evidenced.

BACKGROUND

Local Law 144 was passed by the New York City Council on 10 November 2021 and became law without the Mayor's signature on 13 December 2021 (returned unsigned), with the operational rules to be specified by the Department of Consumer and Worker Protection. The rulemaking process extended through 2022 and into 2023, with two rounds of public comment that surfaced substantial industry and civil-society engagement. The final rules became effective on July 5, 2023, and the law moved from statute to operational regime on that date. The scope is municipal — employers using an AEDT for a hiring or promotion decision affecting a position in New York City — but the practical reach is broad because many employers operating nationally use AEDTs that touch New York City positions.[o]

THE INTERVENTION

The operational requirements have three components. First, an annual independent bias audit by a person or entity that has not used or developed the AEDT, computing selection-rate and impact-ratio metrics by sex, race/ ethnicity, and the intersectional categories the rules specify. Second, publication of a summary of the most recent audit on the employer's website, including the date the audit was conducted, the date the AEDT was first used, and the audit's selection-rate and impact-ratio findings. Third, candidate notice — applicants must be told that an AEDT will be used and given a process to request alternative selection or accommodations. The audit- and-notice structure is the artifact the law produces; the law does not prohibit AEDTs or set bias thresholds for use, and the regulatory theory is disclosure-and-audit rather than substantive standards.[1]

HOW IT WORKED

The independent academic critiques have surfaced two load-bearing limitations. Andrus, Spitzer, Brown, and Xiang's 2021 work on the challenge of procuring demographic data for fairness audits names that many employers do not collect the protected-attribute information the audit metrics require, and the audits that proceed are either limited to attributes the employer happens to have or rely on imputation methods whose accuracy is itself under-evidenced. Wright et al.'s 2024 audit-quality study reviewed published audits across the first cohort of compliant employers and found wide variability — audits ranging from detailed methodological documents to single- paragraph summaries with limited interpretability. The audits are governance artifacts; their information content is uneven across the first cohort.

THE EVIDENCE

The case pairs with Case 85 (OU Analyse) for the governance-objection-dissolved-by-design thread: OU Analyse's deployment surfaced an equity question that the design process resolved structurally; Local Law 144′s audit regime surfaces equity questions structurally through disclosure rather than through a design change. Pair with Case 86 (Gándara community-college predictive equity) for the predictive-equity thread at higher- education scale. Pair with Case 182 (Amazon hiring AI) for the same domain — the audit regime is the governance instrument that, had it been in place, would have applied to an internal recruiting tool of Amazon's described character had it been deployed against NYC candidates. The case is the rare example in the corpus of an intervention at the regulatory scale; whether the audit regime reduces actual disparate impact at the hiring level remains under-evidenced, and the open evaluation question is part of the case. A December 2025 New York State Comptroller audit sharpened that question from the enforcement side: it found the Department of Consumer and Worker Protection's enforcement of the law "ineffective" — flagging only one likely non-compliant disclosure among the 32 it reviewed, where the Comptroller identified at least 17, and finding that roughly 75% of test complaint calls to 311 never reached the agency — moving the open issue from whether disclosure reduces disparate impact to whether the regime is being enforced at all.[2]

WHAT TRANSFERRED

The load-bearing hedges are explicit. The bias-audit-as- governance-artifact intervention is an audit-and-disclosure regime, not a substantive-standards regime; the law does not require employers to achieve any specific impact ratio or to refrain from deploying an AEDT that performs poorly on the audit. Whether the disclosure-and-audit structure reduces

actual disparate impact at the hiring level is an empirical question the published evidence does not yet resolve. The audit-quality variability across the first compliant cohort is itself a finding the case carries — governance artifacts whose quality is uneven across producers are weaker governance instruments than the regulatory theory assumes. The intervention is real and its limits are real; the change-control-and-disclosure-as- governance-artifacts LEO is anchored by the case at the municipal-regulatory scale, and the evaluation arc the audit regime opens is at the start of its evidence development.

> *The audit-and-notice regime is a disclosure-and-audit instrument, not a substantive-standards instrument; whether it reduces actual disparate impact at the hiring level is an empirical question the published evidence does not yet resolve.*
>
> EDITORS' SYNTHESIS OF THE LOCAL LAW 144 RULE TEXT AND THE ANDRUS ET AL. AND WRIGHT ET AL. ACADEMIC CRITIQUES.

REFERENCES

1. New York City Department of Consumer and Worker Protection, Rules Implementing Local Law 144 of 2021 (Automated Employment Decision Tools), effective July 5, 2023.
2. Wright, L., Muenster, R. M., Vecchione, B., Qu, T., Cai, S., Smith, A., Metcalf, J., & Matias, J. N. (2024), "Null Compliance: NYC Local Law 144 and the Challenges of Algorithm Accountability," in Proceedings of FAccT 2024, doi:10.1145/3630106.3658998.
3. Engler, A. (2023), "The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment," Brookings Institution commentary — regulatory-comparative frame for the municipal intervention.
4. Office of the New York State Comptroller (2025), audit of the Department of Consumer and Worker Protection's enforcement of Local Law 144 — finding enforcement "ineffective" (December 2025).

THE LEARNING ENGINEERING LENS

LE INSIGHT

NYC Local Law 144 is the bias-audit-as-governance-artifact intervention at municipal-regulatory scale. The audit-and- notice regime is the first national municipal regulation of algorithmic hiring tools at this scope; the Andrus et al. and Wright et al. critiques name the data-availability and audit-quality limitations the regulatory theory does not provide for. The intervention is real; whether it reduces actual disparate impact at the hiring level is under- evidenced and is the open evaluation question the case carries.

LENS APPROACH

NYC Local Law 144 is the change-control-and-disclosure-as- governance-artifacts case at municipal-regulatory scale (induced 5.4; LENS D5/PT5; LEO-4 and LEO-5). LENS uses it in Domain 5 (Navigating Sociotechnical Constraints) for the audit-as-governance-instrument discipline. Pair with Case 85 (OU Analyse governance-objection-dissolved-by-design), Case 86 (Gándara community-college predictive equity), and Case 182 (Amazon hiring AI). The intervention is real and its limits are real; the disclosure-and-audit structure is not a substantive-standards structure, and the reduction- of-actual-disparate-impact evaluation question is at the start of its evidence development.

DOMAIN	MODES ADDRESSED	LENS COMPETENCY	PROBLEM TYPE	INDUCED	LEO
GOVERNMENT HIRING ALGORITHMS AUDIT AND DISCLOSURE	D G K	D1 D2 D3 D4 **D5** Navigating Sociotechnical Constraints	5	5.4	4 · 5

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Specify the protected-attribute data the audit metrics will require before the audit is commissioned; the Andrus et al. critique names data availability as the precondition the regulatory theory does not provide for, and the data infrastructure has to be built in advance of the audit.
- Choose an independent auditor whose methodology will produce a documentation-detailed audit rather than a single-paragraph summary; the audit-quality variability the Wright et al. study found is itself a deployment choice, and the choice of auditor is where it surfaces.
- Build the candidate-notice and alternative-selection process as part of the deployment, not as a compliance afterthought; the candidate-side governance interaction is the seam at which the disclosure-and-audit structure becomes contestable for the affected person.

IN OPERATION / AFTER

- Carry the load-bearing hedges — disclosure-and-audit regime not substantive-standards regime; reduction of actual disparate impact under-evidenced; audit-quality variability across first cohort — into print without softening; the case's pedagogical value depends on the intervention's limits being preserved.
- Pair in syllabi with Case 85 (OU Analyse) so the governance-objection-dissolved-by-design and governance-objection-surfaced-by-disclosure threads are taught together as complementary intervention forms.
- Use the case as the change-control-and-disclosure-as-governance-artifacts anchor at the municipal-regulatory scale; the curricular target is the discipline of building the data infrastructure and the audit-quality criteria the disclosure regime presupposes.

REFLECTION QUESTIONS

1. Identify a regulated decision domain in your setting in which a disclosure-and-audit regime has been proposed or adopted. What is the protected-attribute data infrastructure the audit metrics will require, and is the infrastructure in place before the regime's effective date?

2. Specify the audit-quality criteria you would apply when commissioning an independent audit. What is the format that distinguishes a documentation-detailed audit from a single-paragraph compliance summary, and what is the decision rule for accepting an auditor's methodology?

3. The Local Law 144 regime is at the start of its evidence development on whether disclosure-and-audit reduces actual disparate impact. Pick a regulatory intervention in your domain and ask: what is the empirical-evaluation arc that would surface the intervention's effect, and what evidence-development infrastructure would the arc require?

WHO BUILDS THIS systems & human-factors engineering · governance, ethics & measurement · knowledge management & data engineering — plus domain experts and a learning engineer to integrate. TOOLS task analysis · design prototyping · governance frameworks · bias & evidence audits · knowledge bases · data pipelines

CASE 205 EDUCATION AT SCALE HEALTHCARE DEFENSE ONGOING

The Discipline We Build Next

T Training Gap K Knowledge & Institutional Memory N Normalization of Deviance G Governance & Trust
H Human-System Interface D Designed Out

IMPACT **Open: the case that has not yet been written. The discipline that this casebook documents is the one practitioners now in training must build.**

IN BRIEF

The cases that precede this one are closed: the investigations are complete, the failures diagnosed, the interventions evaluated. This one is open. It belongs to the practitioners now in training — the students of the LDT program and the LENS specialization, the graduates who will take positions in healthcare, defense, education at scale, and the human-AI frontier, and the institutions they will build. The strongest evidence for the discipline is the cumulative record of what its absence has cost; the strongest argument for its possibility is the record of what its presence has produced. Capability engineering in 2026 sits roughly where INPO sat in 1979 or CRM in 1981 — evidence accumulating, practitioners in training, institutional architecture not yet complete. What the reader does next is what fills in this case.

THE SHIFT

Every preceding case in this book is closed: the investigation is complete, the failure diagnosed or the intervention evaluated. This one is open. It belongs to the practitioners now in training — the students of the LDT program and the LENS specialization, and the institutions they will build across healthcare, defense, education, and the human-AI frontier. What distinguishes this case is not a verdict already reached but a verdict still to be earned, written by the choices its readers make once they leave the page.[0]

WHAT IS EMERGING

What is emerging is a discipline. The argument of this volume is that capability is engineerable — that training, interfaces, evidence systems, and institutions can be designed rather than left to chance — and that the work of designing them is a teachable practice with its own methods and growing body of knowledge. If that argument holds, then the gaps the preceding cases document are not the price of progress but the absence of work that could have been done.[1]

THE CAPABILITY QUESTION

The capability question this case poses is the largest one: can the discipline be built fast enough, and at enough scale, to close the gaps the preceding cases document? The strongest evidence for the discipline is the cumulative record of what its absence has cost; the strongest argument for its possibility is the record of what its presence has produced

— and the case stays open precisely because that question is answered not by argument but by what the next generation of practitioners actually builds.[2]

EARLY EVIDENCE

The early evidence is the success cases themselves. Capability engineering in 2026 sits roughly where INPO sat in 1979 or CRM in 1981 — a discipline whose evidence is accumulating, whose practitioners are being trained, and whose institutional architecture is not yet complete. Those precedents became mature institutions within a generation, which is the most that can yet be claimed here: a comparable trajectory is plausible, but it is not given, and the comparison is an invitation rather than a guarantee.[3]

OPEN PROBLEMS

No specific incident sits behind this case because the work is not done. The institutions of capability engineering that exist today are insufficient to the volume of capability work the world now requires. What the practitioners reading this volume do next — a failure prevented, a success engineered, an institution built — is what fills in this case, and the page is left deliberately unfinished so that the record of the discipline's possibility can still be added to.[4]

> *Capability is engineerable. The institution to engineer it has not yet been built. That is the work.*
>
> LDT / LENS PROGRAM STATEMENT, JOHNS HOPKINS SCHOOL OF EDUCATION

REFERENCES

1. LDT / LENS program statement, Johns Hopkins School of Education — the program and its commitments.
2. Goodell & Kolodner (2022), Learning Engineering Toolkit — the discipline's methods.
3. The preceding cases of this volume — the cumulative record of cost and of possibility.
4. INPO (Case 175) and CRM/CAST (Case 117) — institution-building precedents (1979, 1981).
5. IEEE ICICLE Learning Engineering Body of Knowledge (LEBoK); the introduction to this volume.

THE LEARNING ENGINEERING LENS

LE INSIGHT

Every case in this book is closed. This one is open. The discipline LENS represents is the discipline that closes cases that are now open. The institutions of capability engineering that exist in 2026 are insufficient to the volume of capability work the contemporary world requires. The graduates of this program are the people who will build the institutions that are insufficient now.

LENS APPROACH

LENS is organized to produce the practitioners who can do this work. The curriculum's commitments — capability as a system parameter, evidence as a deliverable, ethics as design, implementation as a thread — are the operational form of the argument the preceding cases evidence. Every studio project (LEN 10) is a small attempt at this case.

DOMAIN

EDUCATION AT SCALE · HEALTHCARE · DEFENSE

MODES AT STAKE

T K N G H D

APPROACHES TO CONSIDER

DURING DEVELOPMENT

- Treat capability as a system parameter to be specified and engineered from the outset of any deployment, not left to chance and diagnosed only after failure.
- Build the evidence system as a deliverable alongside the capability, so the discipline can show what its presence produces rather than only argue for it.
- Design the institution that will carry the practice forward, taking INPO and CRM as precedents for how a discipline matures within a generation.

IN OPERATION / AFTER

- Monitor whether the discipline is in fact closing the gaps the preceding cases document, treating the open case as a standing measure of progress.
- Sustain the institutional architecture once built, since the precedents show maturity is reached only by carrying the work across a generation.
- Keep the cumulative record of cost and of possibility current, so each failure prevented or success engineered is added to the evidence the discipline rests on.

REFLECTION QUESTIONS

1. What is the case you are positioned to add to this dataset in the next ten years — a failure prevented, a success engineered, an institution built?
2. If this book is read by the practitioners who will write its next case, what should that case look like?
3. Capability engineering in 2026 is compared here to where INPO stood in 1979 — a plausible trajectory, not a guaranteed one. What would have to be true a generation from now for that comparison to hold, and what is the first institution you would build to make it so?

WHO BUILDS THIS learning science & instructional design · knowledge management & data engineering · safety science & organizational analysis · governance, ethics & measurement · human factors & interaction design · systems & human-factors engineering — plus domain experts and a learning engineer to integrate. TOOLS scenario simulation · competency models · knowledge bases · data pipelines · incident analysis · safety dashboards · governance frameworks · bias & evidence audits · usability testing · cognitive walkthroughs · task analysis · design prototyping

DOMAIN INDEX

The cases by domain.

The matrix at the front lists every case in number order. This index reorganises the cases by primary domain — a reader following healthcare, defense, education, or any other thread can find the relevant cases together. Each case appears once, under its primary domain. Where a section opens with a callout, it names the dataset's standout failure and, where one exists in that domain, its standout success.

DEFENSE

STANDOUT FAILURE	Case 5	**Operation Eagle Claw (1980).** Cross-organizational capability that existed on no service's deliverable list; the mission was improvised across the Army, Navy, Marines, Air Force, and CIA. Produced the Holloway Commission, USSOCOM, and Goldwater-Nichols.
STANDOUT SUCCESS	Case 18	**U.S. Nuclear Navy / Rickover Training Model (1954–).** The longest continuous capability-engineering program in any high-consequence domain. Sixty-plus years of zero reactor accidents through training treated as a system parameter.

AVIATION

STANDOUT FAILURE **Case 7** **Boeing 737 MAX / MCAS (2018–19).** A single-string angle-of- attack sensor, a flight-control law that assumed pilot mental models from prior models, and a certification chain that did not knit the pieces together. 346 lives.

STANDOUT SUCCESS **Case 70** **Crew Resource Management & CAST (1981–).** First-officer authority engineered into the cockpit, a continuous learning architecture across the industry, and the safety record that followed.

HEALTHCARE

STANDOUT FAILURE **Case 61**

EHR / CPOE Implementation (2005–). Roughly $30B federal investment in systems designed for billing and administration, deployed against clinical workflow. Usability is now itself a patient-safety variable at scale.

STANDOUT SUCCESS **Case 109**

WHO Surgical Safety Checklist (2008–). One page, nineteen items, paired with a required pause. Halved surgical mortality in the multi-site pilot; the smallest effective capability intervention in the dataset.

EDUCATION AT SCALE

STANDOUT FAILURE — Case 149

Summit Learning / Personalized Learning Rollout (2014–19). A defensible pedagogical platform, well funded, deployed across 380 schools — and pulled by parent revolt because the governance architecture (consent, evidence, exit) was never engineered alongside it.

STANDOUT SUCCESS — Case 110

Georgia State University Predictive Analytics (2012–). 800 risk factors per student, advising triggered by early warning signs, equity as a primary constraint. Six-year graduation 32% → 54%; equity gaps eliminated.

ENERGY

STANDOUT FAILURE — Case 69

Three Mile Island (1979). Training matched the design-basis events the engineers had imagined; the accident was not one of them. The control-room interface confused command signal with valve state. Catalysed industry-wide reform.

STANDOUT SUCCESS — Case 172

INPO and the Nuclear Academy (1979–). The industry built the institution TMI revealed it needed. No INES-level event at U.S. commercial reactors since.

GOVERNMENT

STANDOUT FAILURE — Case 162

Australia Robodebt (2016–20). An income-averaging algorithm that reversed the burden of proof onto welfare recipients. The Royal Commission found "venality, incompetence and cowardice"; the scheme was linked to deaths, including suicide.

NOTE — Case 111

No paired-intervention success in the government dataset. The reforms that follow these cases — Royal Commissions, statutory restructures — are documented in the narratives but did not produce a LENS-style success case in time for this volume.

INDUSTRIAL

STANDOUT FAILURE **Case 147** **Bhopal (1984).** Training, knowledge, normalisation of deviance, and governance all collapsed in the same plant. Roughly 4,000 immediate deaths; chronic harms continue. The cautionary case for outsourced operational responsibility.

STANDOUT SUCCESS **Case 73** **Toyota Production System / Andon Cord (1950s–).** Any operator can stop the line. The smallest authority-architecture intervention in the dataset, and the most durably copied.

TECHNOLOGY

STANDOUT FAILURE **Case 66** **UK Post Office Horizon Scandal (1999–2015).** Institutional refusal to believe sub-postmasters' reports of computer error, sustained over fifteen years. Hundreds of wrongful convictions; ruined lives and at least four suicides — the longest-running operator-feedback failure in the dataset.

NOTE **Case 132** The technology dataset's success story is filed under Healthcare: AlphaFold (Case 36), the contemporary worked example of computational capability-engineering done as a discipline.

AI IN EDUCATION

K-12 EDUCATION

K-12 MATHEMATICS

K-12 SCIENCE

AIR TRAFFIC MANAGEMENT

ANESTHESIOLOGY

AUTONOMOUS SYSTEMS

STANDOUT FAILURE **Case 62** — **Uber ATG / Tempe Fatality (2018).** The safety driver was nominally responsible for what the system was designed to do without her attention — the classic vigilance contradiction at the heart of partial-autonomy operator roles.

NOTE **Case 194** — No paired-intervention success in the autonomous-systems dataset. The capability engineering this category needs is still being built — see Case 205 (**The Discipline We Build Next**) for the open question this volume closes on.

AUTONOMOUS VEHICLES

AVIATION INFRASTRUCTURE

AVIATION SAFETY

CLINICAL CARE

EMERGENCY MANAGEMENT

FINANCE

FINANCIAL SERVICES

GENERAL AVIATION

GLOBAL HEALTH

GOVERNMENT

GOVERNMENT/WELFARE (NETHERLANDS)

HEALTH PROFESSIONS EDUCATION

HUMANITARIAN RELIEF

IMPLEMENTATION SCIENCE

LEARNING ANALYTICS

MEDICAL DEVICES

MEDICAL EDUCATION

MEDICAL-DEVICE REGULATION

MULTIMODAL LEARNING ANALYTICS

NAVAL ENGINEERING

NEUROSCIENCE

SPACE / AEROSPACE

STANDOUT FAILURE	Case 140	**Challenger & Columbia (1986 / 2003).** Normalisation of deviance across two decades and two crews, captured in the Rogers and CAIB reports as a culture that accepted flying with flaws when problems were defined as normal and routine.
NOTE	Case 18	No paired-intervention success in the aerospace dataset. See Defense (Rickover, Case 137) for the safety-engineering counterpart that aerospace did not build.

SURGERY

TECHNOLOGY

TUTORING

WORKFORCE DEV

About this index. Each case is filed under its primary domain, so every case appears exactly once. Cases that span several domains (a healthcare-AI case is also a technology case, a defense-training case is also an

> education case) are cross-referenced in their narratives. The standout failures and successes are editorial picks — the case in that domain the editors return to most often in studio.

LENS COURSE INDEX

The cases by LENS course.

Each case names the LEN courses it serves as a worked example for. This index inverts that mapping: for every course in the specialization, the cases that inform it, in case-number order. A case appears under every course it supports, so many appear more than once.

LEN 1

36 cases

LEN 2

53 cases

LEN 3

47 cases

LEN 4

68 cases

LEN 5 84 cases

LEN 6

17 cases

LEN 7

102 cases

LEN 8

90 cases

LEN 9

30 cases

REFERENCES

Works cited.

Numbered references are cited in the Introduction. Per-case sources appear on each spread. The broader reading list at the end of this section collects foundational works the LENS curriculum draws on.

CITED IN THE INTRODUCTION

1. Makary, M. A. & Daniel, M. (2016). Medical error — the third leading cause of death in the US. BMJ 353: i2139. doi:10.1136/bmj.i2139.
2. Morris, Z. S., Wooding, S. & Grant, J. (2011). The answer is 17 years, what is the question: understanding time lags in translational research. Journal of the Royal Society of Medicine 104(12): 510–520. doi:10.1258/jrsm.2011.110180.
3. Balas, E. A. & Boren, S. A. (2000). Managing clinical knowledge for health care improvement. Yearbook of Medical Informatics, 65–70.
4. U.S. Navy (2017). Comprehensive Review of Recent Surface Force Incidents; Strategic Readiness Review. Department of the Navy.
5. U.S. Government Accountability Office (2023). F-35 Sustainment: DOD Faces Several Uncertainties and Risks as it Defines Future Costs and Plans. GAO-23-105341. gao.gov.
6. Bulger, M., McCormick, P. & Pitcan, M. (2017). The Legacy of inBloom. Data & Society Research Institute. datasociety.net.
7. Office of Qualifications and Examinations Regulation, UK (2020). Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: interim report. Ofqual; see also Adam, D. (2020), MIT Technology Review.
8. Helmreich, R. L., Merritt, A. C. & Wilhelm, J. A. (1999). The evolution of Crew Resource Management training in commercial aviation. International Journal of Aviation Psychology 9(1): 19–32. doi:10.1207/s15327108ijap0901_2.
9. Commercial Aviation Safety Team / Federal Aviation Administration (2016). Safety Enhancements and Outcomes. CAST/FAA.
10. Federal Aviation Administration (2017). Out Front on Airline Safety: Two Decades of Continuous Evolution. FAA Newsroom.
11. Rees, J. V. (1994). Hostages of Each Other: The Transformation of Nuclear Safety since Three Mile Island. University of Chicago Press. ISBN 978-0226706887.
12. Pronovost, P., Needham, D., Berenholtz, S. et al. (2006). An intervention to decrease catheter-related bloodstream infections in the ICU. New England Journal of Medicine 355(26): 2725–2732. doi:10.1056/NEJMoa061115.
13. Haynes, A. B., Weiser, T. G., Berry, W. R. et al. (2009). A surgical safety checklist to reduce morbidity and mortality in a global population. New England Journal of Medicine 360(5): 491–499. doi:10.1056/NEJMsa0810119.
14. Agency for Healthcare Research and Quality (2023). TeamSTEPPS 3.0 Curriculum. AHRQ Publication No. 23-0034. ahrq.gov.
15. Sawyer, R. K. (ed.) (2014). The Cambridge Handbook of the Learning Sciences (2nd ed.). Cambridge University Press. ISBN 978-1107033252.
16. Vicente, K. J. (1999). Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work. Lawrence Erlbaum. ISBN 978-0805823974.
17. Wickens, C. D., Hollands, J. G., Banbury, S. & Parasuraman, R. (2021). Engineering Psychology and Human Performance (5th ed.). Routledge. doi:10.4324/9781003177616.
18. Fixsen, D. L., Naoom, S. F., Blasé, K. A., Friedman, R. M. & Wallace, F. (2005). Implementation Research: A Synthesis of the Literature. National Implementation Research Network. See also Aarons, G. A., Hurlburt, M. & Horwitz, S. M. (2011). Advancing a conceptual model of evidence-based practice implementation in public service sectors. Administration and Policy in Mental Health 38: 4–23. doi:10.1007/s10488-010-0327-7.
19. Reason, J. (1990). Human Error. Cambridge University Press, doi:10.1017/CBO9781139062367. Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. Safety Science 27: 183–213, doi:10.1016/S0925-7535(97)00052-0. Leveson, N. (2011). Engineering a Safer World: Systems Thinking Applied to Safety. MIT Press, doi:10.7551/mitpress/8179.001.0001.

20. Weick, K. E. & Sutcliffe, K. M. (2015). Managing the Unexpected: Sustained Performance in a Complex World (3rd ed.). Wiley. See also Roberts, K. H. (1990). Some characteristics of one type of high reliability organization. Organization Science 1(2): 160–176.
21. Saxberg, B. (2017). Learning engineering: the art of applying learning science at scale. In Industry Connections Industry Consortium on Learning Engineering. See also Goodell, J. & Kolodner, J. L. (eds.) (2022). Learning Engineering Toolkit. Routledge.
22. IEEE ICICLE (2022). Learning Engineering Body of Knowledge (LEBoK). lebok.wiki.
23. MICrONS Consortium (2025). Functional connectomics spanning multiple areas of mouse visual cortex. Nature 640: 167–179. doi:10.1038/s41586-025-08790-w. See also Hider, R. et al. (2022), BossDB, Frontiers in Neuroinformatics; Xenes, D. et al. (2022), NeuVue, bioRxiv 2022.07.18.500521; Gray-Roncal, W. (in press, 2026), "Calibrate the Scientists, Not Just the Microscopes," BRAIN CONNECTS commentary.
24. Cervantes, M., Floryanzia, S., Sharp, J., Johnson, E. C. & Gray-Roncal, W. R. (2023). Empowering trailblazers toward scalable, systematized, research-based workforce development. 2023 ASEE Annual Conference & Exposition, Baltimore, MD. peer.asee.org. Documents the CIRCUIT model; Gray-Roncal is the senior author. See also Gray-Roncal, W. et al. (2025), MERIT: Mentoring Exceptional Researchers to Innovate and Thrive — A Structured, Scalable Research Training Framework for STEM Talent Development. JHU/APL whitepaper (the generalization of CIRCUIT). COMPASS — the layer of targeted mentor interventions that operates MERIT inside specific institutional settings, designed to help students uncover tacit and explicit knowledge gaps. COMPASS-Core: Curriculum Overview and Workshop Summaries (JHU/APL program documentation) is one well-documented instance, a workshop set grounded in the CCR framework and explicitly addressing the "hidden curriculum" of research practice.
25. TITAN — DARPA Artificial Intelligence Exploration program (JHU/APL pilot, c. 2020 – 2022). Program documentation on file with the LENS partner team at the JHU/APL AI Pathfinding Lab.
26. PISTA (2025). Perceptual Inference System for Tactical Agentic AI. JHU/APL AI Pathfinding Lab, Summer 2025 Research Incubator, sponsored by DTRA JSTO. Cohort poster and program documentation available from the LENS partner team.
27. Precision-medicine community-detection methods — preprint at medRxiv 2024.06.20.24309267.

FOUNDATIONS · THE DISCIPLINE DRAWS ON

Learning sciences and learning engineering

• Bransford, J. D., Brown, A. L. & Cocking, R. R. (eds.) (2000). How People Learn: Brain, Mind, Experience, and School. National Academy Press. doi:10.17226/9853.
• Sawyer, R. K. (ed.) (2014). The Cambridge Handbook of the Learning Sciences (2nd ed.).
• Koedinger, K. R., Corbett, A. T. & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework. Cognitive Science 36(5): 757–798. doi:10.1111/j.1551-6709.2012.01245.x.
• Sweller, J., Ayres, P. & Kalyuga, S. (2011). Cognitive Load Theory. Springer. doi:10.1007/978-1-4419-8126-4.
• Soderstrom, N. C. & Bjork, R. A. (2015). Learning versus performance: an integrative review. Perspectives on Psychological Science 10(2): 176–199. doi:10.1177/1745691615569000.
• Goodell, J. & Kolodner, J. L. (eds.) (2022). Learning Engineering Toolkit: Evidence-Based Practices from the Learning Sciences, Instructional Design, and Beyond. Routledge. doi:10.4324/9781003276579.

Human factors, cognitive engineering, and systems safety

• Vicente, K. J. (1999). Cognitive Work Analysis. Lawrence Erlbaum.
• Norman, D. A. (2013). The Design of Everyday Things (rev. ed.). Basic Books. ISBN 978-0465050659.
• Hollnagel, E. & Woods, D. D. (2005). Joint Cognitive Systems: Foundations of Cognitive Systems Engineering. CRC Press. doi:10.1201/9781420038194.
• Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors 37(1): 32–64. doi:10.1518/001872095779049543.
• Perrow, C. (1984). Normal Accidents: Living with High-Risk Technologies. Basic Books. ISBN 978-0691004129 (1999 Princeton reprint).
• Leveson, N. (2011). Engineering a Safer World: Systems Thinking Applied to Safety. MIT Press. doi:10.7551/mitpress/8179.001.0001.
• Reason, J. (1997). Managing the Risks of Organizational Accidents. Ashgate. ISBN 978-1840141047.
• Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. Safety Science 27: 183–213. doi:10.1016/S0925-7535(97)00052-0.
• Dekker, S. (2014). The Field Guide to Understanding 'Human Error' (3rd ed.). Ashgate. ISBN 978-1472439055.

High-reliability organizing and team performance

• Weick, K. E. & Sutcliffe, K. M. (2015). Managing the Unexpected (3rd ed.). Wiley. ISBN 978-1118862414.
• Roberts, K. H. (1990). Some characteristics of one type of high reliability organization. Organization Science 1(2): 160–176. doi:10.1287/orsc.1.2.160.
• Edmondson, A. C. (2018). The Fearless Organization: Creating Psychological Safety in the Workplace for Learning, Innovation, and Growth. Wiley. ISBN 978-1119477242.

• Salas, E., Wilson, K. A., Burke, C. S. & Wightman, D. C. (2006). Does Crew Resource Management training work? An update, an extension, and some critical needs. Human Factors 48(2): 392–412. doi:10.1518/001872006777724444.
• Klein, G. (1998). Sources of Power: How People Make Decisions. MIT Press. ISBN 978-0262112277.
• Kanki, B. G., Helmreich, R. L. & Anca, J. (eds.) (2010). Crew Resource Management (2nd ed.). Academic Press. ISBN 978-0123749468.

Implementation science and organizational learning

• Fixsen, D. L., Naoom, S. F., Blasé, K. A., Friedman, R. M. & Wallace, F. (2005). Implementation Research: A Synthesis of the Literature. NIRN.
• Aarons, G. A., Hurlburt, M. & Horwitz, S. M. (2011). Advancing a conceptual model of evidence-based practice implementation in public service sectors. Administration and Policy in Mental Health 38: 4–23. doi:10.1007/s10488-010-0327-7.
• Damschroder, L. J. et al. (2009). Fostering implementation of health services research findings into practice: a consolidated framework. Implementation Science 4: 50. doi:10.1186/1748-5908-4-50.
• Argyris, C. & Schön, D. A. (1996). Organizational Learning II: Theory, Method, and Practice. Addison-Wesley. ISBN 978-0201629835.
• Senge, P. M. (2006). The Fifth Discipline: The Art and Practice of the Learning Organization (rev. ed.). Doubleday. ISBN 978-0385517256.
• Vaughan, D. (1996). The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA. University of Chicago Press. ISBN 978-0226851761.

Ethics, governance, and equity in sociotechnical systems

• Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's. ISBN 978-1250074317.
• Friedman, B. & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems 14(3): 330–347. doi:10.1145/230538.230561.
• O'Neil, C. (2016). Weapons of Math Destruction. Crown. ISBN 978-0553418811.
• Selbst, A. D. et al. (2019). Fairness and abstraction in sociotechnical systems. In FAT*'19. doi:10.1145/3287560.3287598.
• Citron, D. K. (2008). Technological due process. Washington University Law Review 85: 1249–1313. SSRN.
• Selwyn, N. (2014). Distrusting Educational Technology. Routledge. ISBN 978-1138021068.

Human-AI teaming and automation

• Parasuraman, R. & Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. Human Factors 52(3): 381–410. doi:10.1177/0018720810376055.
• Bainbridge, L. (1983). Ironies of automation. Automatica 19(6): 775–779. doi:10.1016/0005-1098(83)90046-8.
• Sarter, N. B. & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. Human Factors 37(1): 5–19. doi:10.1518/001872095779049516.
• Cummings, M. L. (2017). Automation bias in intelligent time-critical decision support systems. In Decision Making in Aviation.
• Casner, S. M. & Hutchins, E. (2019). What do we tell the drivers? Toward minimum driver-training standards for partially-automated cars. Journal of Cognitive Engineering and Decision Making 13(2): 55–66. doi:10.1177/1555343419830901.

Patient safety and clinical improvement

• Kohn, L. T., Corrigan, J. M. & Donaldson, M. S. (eds.) (1999). To Err Is Human: Building a Safer Health System. Institute of Medicine. doi:10.17226/9728.
• Gawande, A. (2009). The Checklist Manifesto: How to Get Things Right. Metropolitan. ISBN 978-0805091748.
• Pronovost, P. & Vohr, E. (2010). Safe Patients, Smart Hospitals. Hudson Street Press. ISBN 978-1594630644.
• Wachter, R. (2015). The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age. McGraw-Hill. ISBN 978-0071849463.
• Bates, D. W., Levine, D. M., Salmasian, H. et al. (2023). The safety of inpatient health care. New England Journal of Medicine 388(2): 142–153. doi:10.1056/NEJMsa2206117.

Each case spread in the body of the book includes its own primary sources. Sources cited only within case spreads are not repeated here; this list collects the works the discipline as a whole draws on, together with the references used in the introductory essay.

APPENDIX · LEOS AND COURSE COVERAGE

The competencies, the LEOs, and the courses.

COMPETENCY → CONCENTRATION LEARNING OUTCOME

The five capabilities the cases turn on are named, in the curriculum, as five LENS Educational Objectives (LEOs) — the concentration-level objectives, held at the program level:

Systems Analysis	LEO-1 · Systems Thinking and Analysis
Iterative Development	LEO-2 · Learning Engineering Design and Implementation
Human-System Collaboration	LEO-3 · Human-System Collaboration and Adaptive Systems
Test and Evaluation	LEO-4 · Data, Measurement, and Evaluation
Navigating Sociotechnical Constraints	LEO-5 · Context and Domain Fluency

THE LEN COURSES · CASES PER COURSE

Each case names the LEN courses it serves as a worked example for, and most cases serve several. The counts below are therefore not a partition of the 202 cases — they sum to well more than 202, because a single case typically appears under three or four courses.

LEN 1	Principles of learning engineering	36
LEN 2	Human-AI teaming	54
LEN 3	Systems integration	49

LEN 4	Evidence and measurement	67
LEN 5	Capability analysis	86
LEN 6	Applied problem-solving	18
LEN 7	Bias and governance	114
LEN 8	Knowledge transfer	98
LEN 9	Computational methods	30
LEN 10	Capstone studio	19

WHAT THE DISTRIBUTION SHOWS

The distribution is itself diagnostic. The strongly covered topics are the ones the existing case literature most readily supports: bias and governance (LEN 7) and knowledge transfer (LEN 8) lead, with capability analysis (LEN 5) and evidence and measurement (LEN 4) close behind. Across 202 cases the corpus is saturated with examples of what goes wrong when ethics, governance, evidence, and institutional learning are not engineered — and, in the success cases, what it takes to engineer them.

The thinner coverage is equally informative. Computational methods (LEN 9) has the failure modes of algorithmic systems well represented but the computational tools for capability instrumentation under-represented. The capstone (LEN 10) maps to fewer cases than the number of studio projects each cohort undertakes. Systems integration (LEN 3) and applied problem-solving (LEN 6) required focused tagging to reach defensible coverage, surfacing that operational systems-engineering content is under-represented in the published case literature relative to its curricular centrality, and that stakeholder-analysis and problem-framing have not historically been packaged as case material by other disciplines. These are pedagogical gaps the LENS studio sequence is organized to close — not by adding more failure cases, but by having students produce the missing artifacts in real operational settings and entering them into the dataset for the next cohort.

WHAT LENS OFFERS

The program behind the casebook.

LENS — **Learning Engineering for Next-Generation Systems** — is a specialization within the Johns Hopkins University School of Education's **Learning Design & Technology** program. It operationalizes the casebook's central claim — that capability is a designable, measurable, engineerable property of a complex system — for high-consequence work in healthcare, defense, and education. It is housed in a School of Education because the binding constraint is rarely the engineering, which is mature; it is the learning sciences, educational measurement, implementation science, and equity-and-governance machinery that decide whether a capability intervention actually scales.

THE FIVE PILLARS

LENS stands on five institutional pillars, load-bearing together rather than singly. **Mission Literacy** — reading what an operational setting actually requires and translating it into design and measurement decisions a serious reviewer would accept. **JHU Ecosystem** — the School of Education's centers, doctoral programs, and faculty in learning sciences, measurement, and design-based research, alongside university practice partners (the Armstrong Institute for Patient Safety and Quality; the Bloomberg School's implementation-science tradition; Whiting School engineering and cognitive science; the Berman Institute's governance and ethics work). **Intersectional Expertise** — the distinguishing pillar: holding the engineering, the learning, the measurement, the equity, and the operational context in one conversation. **Capability Focus** — keeping every artifact, measurement, and iteration accountable to what the system can actually do. **Flywheel Iteration** — the Practice Flywheel (**Identify → Activate → Prototype → Analyze → Transition**) that turns each cohort, case, and pilot into the starting condition of the next.

THE RECORD AT JOHNS HOPKINS

The placement is not a default. The School of Education has studied the learner, designed the assessment, and engineered the intervention since before learning engineering had its current name. The Center for Talented Youth (1979) holds one of the longest continuous evidence bases in U.S. education on the identification and accelerated instruction of advanced learners; the Center for Research and Reform in Education has, since 2004, run large-scale evaluations against rigorous evidence standards (ESSA tiers, What Works Clearinghouse) and publishes the Best Evidence Encyclopedia. The Learning Design & Technology program inherits that environment: for two decades its M.Ed. and EdD tracks have graduated designers and evaluators of learning systems whose pedagogical core is design-based research. LENS formalizes that orientation as a specialization rather than departing from it.

THE PILOTS

The claim that capability is engineerable rests on a documented record of learning-engineering pilots run at the Johns Hopkins Applied Physics Laboratory before LENS enrolled its first cohort. **CIRCUIT** was built as the workforce solution to the IARPA MICrONS connectomics program; the cohorts it trained contributed materially to the MICrONS reconstructions now published in Nature, with student development a deliberate byproduct of the mission result. The model was generalized into the **MERIT** framework and the **COMPASS** mentor-intervention layer, documented in the peer-reviewed engineering-education literature. The **PISTA** incubator (DTRA, Summer 2025) compressed the same cycle into ten weeks: nine student fellows shipped a working CBRNE decision-support prototype to agency leadership through three full redesigns. No single pilot settles the proposition; the cumulative record shows the program is itself the institutional flywheel its case studies argue every high-consequence domain requires.

ABOUT THE EDITORS

The editors.

The casebook's central argument — that capability engineering is the applied form of the educational sciences, taught from the School of Education and informed by operational practice — is embodied in the two editors. Each has spent a career working in one half of that pairing and a substantial portion of it in the other. Each contributes to the five pillars on which LENS rests: Mission Literacy · JHU Ecosystem · Intersectional Expertise · Capability Focus · Flywheel Iteration.

William Gray-Roncal, Ph.D.

APPLIED RESEARCH · LEARNING ENGINEERING · STEM WORKFORCE

William Gray-Roncal, Ph.D., is Principal Research Engineer at the Johns Hopkins University Applied Physics Laboratory, where he applies systems engineering methods to challenges spanning defense, space, biomedical, and learning systems. His research — sponsored by DARPA, IARPA, NIH, and DoD — bridges computational neuroscience, artificial intelligence, and large-scale data infrastructure. He developed the College Prep and CIRCUIT workforce-development models, through which he has mentored nearly 500 students, and his current work develops quantitative measures of human cognitive performance in operational contexts. He leads the development of the LENS (Learning Engineering for Next-Generation Systems) specialization within the Johns Hopkins Master of Education in Learning Design and Technology, the program directed by James Diamond at the School of Education. Gray-Roncal holds a Ph.D. in Computer Science from Johns Hopkins, an M.S. in Electrical Engineering from USC, and a B.S. in Electrical Engineering from Vanderbilt.

James Diamond, Ph.D.

LEARNING SCIENCES · EDUCATIONAL DESIGN · EVALUATION

James Diamond, Ph.D., is Assistant Professor in the Johns Hopkins University School of Education and Program Director of the Master of Education in Learning Design and Technology, where he applies design-based research and learning-sciences methods to the design and evaluation of technology-mediated learning environments. His research — supported by the National Science Foundation, the Institute of Education Sciences, the Bill & Melinda Gates Foundation, the MacArthur Foundation's HASTAC Digital Media and Learning Initiative, the Robin Hood Learning + Technology Fund, the Wellcome Trust, and the UK Economic and Social Research Council — bridges digital games and simulations for learning, digital badges and micro-credentialing, classroom technology integration, and disciplinary literacy in history, social studies, civics, and STEM. Before joining Johns Hopkins, he was Senior Research Associate at the Education Development Center's Center for Children and Technology in New York City, where he led projects including Zoom In and the Who Built America? Teacher Mastery Badge System, contributed to Mission US and Possible Worlds, and taught at New York University as an adjunct. As Program Director, he oversees the LDT program within which the LENS (Learning Engineering for Next-Generation Systems) specialization is offered. Diamond holds a Ph.D. in Educational Communication and Technology from New York University, an Ed.M. in Educational Technology from Boston University, and a B.A. in History from Boston University.

These biographies reference careers documented in publications, program reports, faculty pages, and the curricular materials of the LDT program and the LENS specialization.