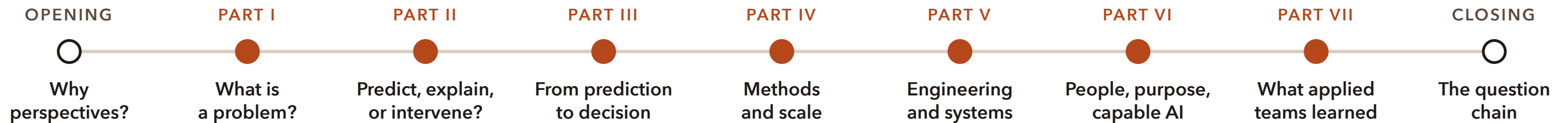




APPLIED AND DEPLOYED · DATA · LEARNING · INTELLIGENT SYSTEMS

Perspectives on Problems

What the pioneers saw, and what still holds in 2026



What will this session do?

AIM

For seventy years, researchers and practitioners have argued about what a problem is and how data and machines should help solve it. This session takes their strongest claims, tests each one against 2026 evidence and deployed systems, and keeps what still works as rules for practice.

WHAT YOU WILL LEARN

1. How pioneers classified problems: ill structured, wicked, messes; predict, explain, or intervene.
2. How to test a classic claim against current evidence, and give a verdict with its limits.
3. How to tell problems where prediction is enough from problems that need causal reasoning.
4. Recurring failures of deployed systems: proxy labels, selective labels, feedback loops, leakage, underspecification, hidden dependencies.
5. Fourteen questions to examine a new problem before you build.

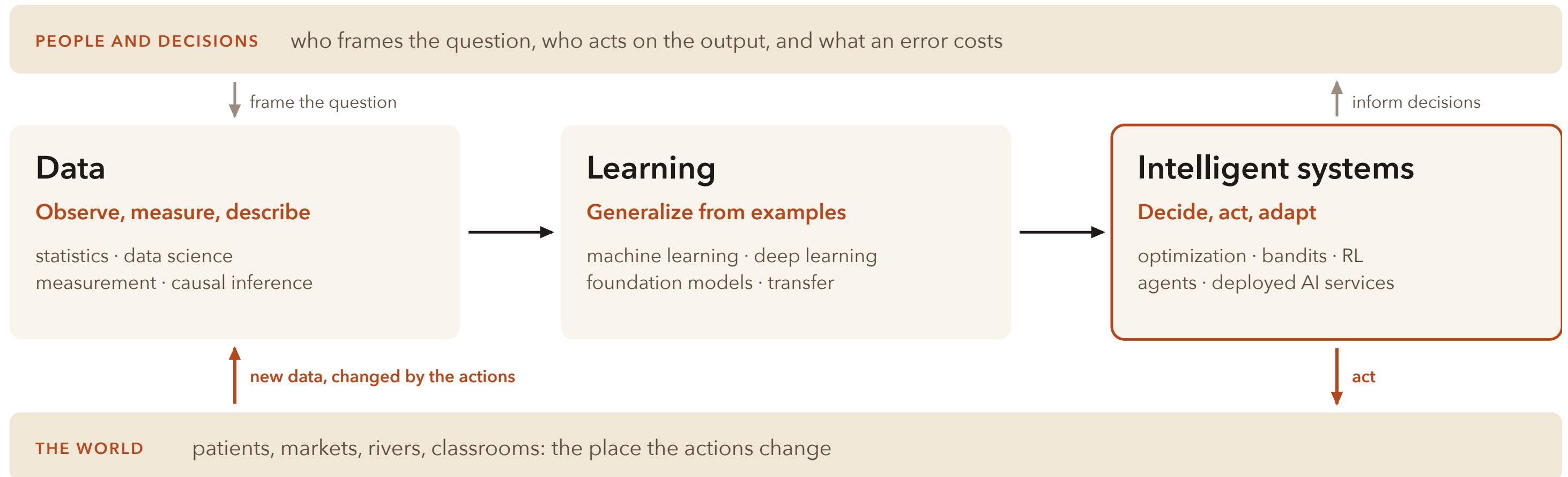
OPENING QUESTION · 2 MINUTES, ON PAPER

Write one belief about solving problems with data and AI that you expect to still hold in 2036.

WHAT WE HIGHLIGHT

- The question comes before the method.
- The label you choose is the policy you set.
- Prediction is not intervention.
- Most of a deployed system is not the model.
- Purpose and accountability stay with people.

What do we mean by data, learning, and intelligent systems?



Data, learning, and intelligent systems are data, models, people, and decisions working together to solve a problem. Their methods range from statistics and machine learning to foundation models, optimization, and simulation.

A WARNING SERVICE

Flood Hub: river gauges, one learned model, warnings in more than 80 countries.

A HOSPITAL ALERT

Epic Sepsis Model: records, a risk score, and clinicians who act on alerts.

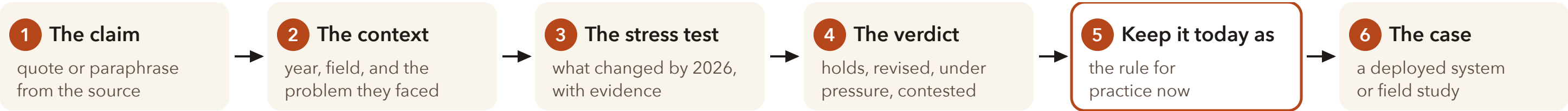
A FEED

YouTube: viewing logs, a ranking model, and people whose choices become the next data.

A PUBLIC DECISION

Ofqual 2020: past results, a grading rule, and students' futures.

How will we test seventy years of ideas?



PERSPECTIVES IN THIS SESSION, BY ERA · COLOUR = 2026 VERDICT

Holds

Holds with revision

Under pressure

Contested

1950s-1970s

Shannon	1952
Meehl	1954
Wiener	1960
Licklider	1960
Engelbart	1962
Tukey	1962
Simon	1962-73
Rittel & Webber	1973
Box	1976
Ackoff	1979

1980s-2000s

Schön	1983
Vapnik	1995-2006
Breiman	2001
Hand	2006
Gigerenzer & Brighton	2009
Halevy, Norvig, Pereira	2009

2010s

Shmueli	2010
Kleinberg et al.	2015
Sculley et al.	2015
Athey	2017
Kleinberg et al.	2018
Booking.com (Bernardi et al.)	2019
Hernán, Hsu, Healy	2019
Jordan	2019
Obermeyer et al.	2019
Passi & Barocas	2019
Pearl	2019
Russell	2019
Sutton	2019

2020s

Perdomo et al.	2020
Geirhos et al.	2020
D'Amour et al.	2022
M5 (Makridakis et al.)	2022
Raji et al.	2022
Kapoor & Narayanan	2023
Kambhampati et al.	2024
Narayanan & Kapoor	2025
Silver & Sutton	2025

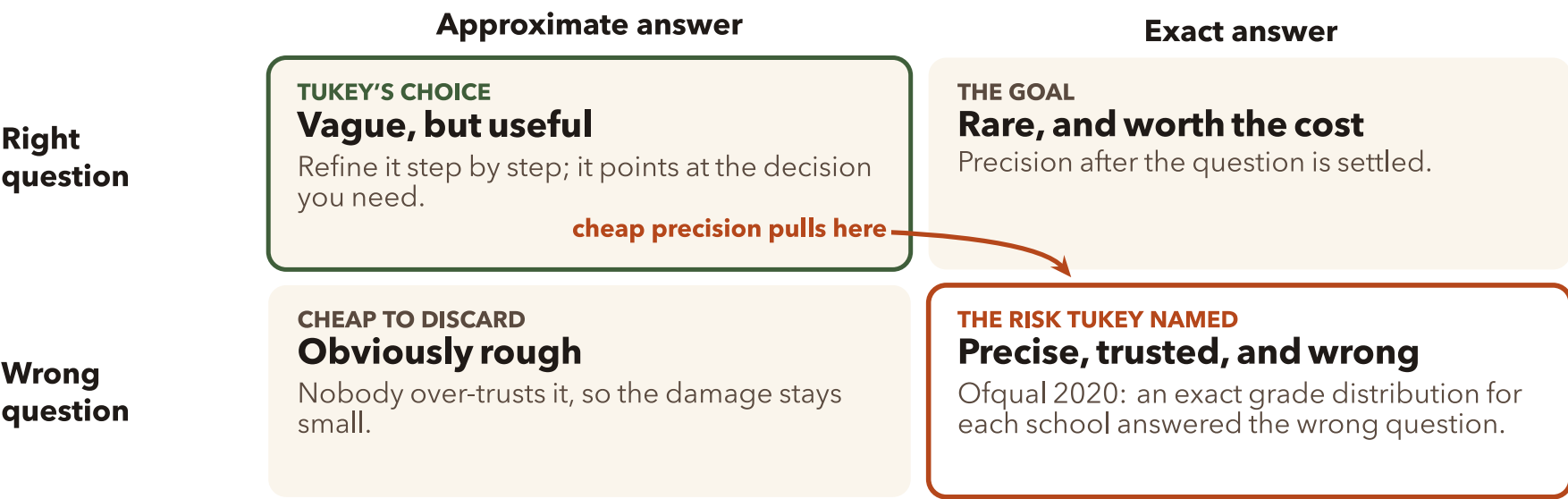
Is an exact answer to the wrong question still the biggest risk?

Tukey’s warning is about the question, not the method: precision cannot rescue a question that was wrong from the start.

“Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.”

John W. Tukey, “The Future of Data Analysis”, Annals of Mathematical Statistics, 1962, pp. 13–14

Context. A statistician arguing that data analysis is a science of judgment, not only of proof. He calls this “the most important maxim for data analysis to heed.”



The rows are the question you ask, the columns the answer you get. Cheap precision pushes work toward the bottom-right corner, where errors are trusted.

2026 STRESS TEST

Holds, and stronger

Precise numbers are now cheap: leaderboards, benchmark scores, ratings from LLM judges. Private testing of many model variants before release skews public leaderboards (Singh et al. 2025).

Cheap precision makes the wrong question more common, not less.

CASE

Ofqual A-level grading, England, 2020

Exams were cancelled. An algorithm moderated teacher grades toward each school’s 2017–19 results. **39.1%** of grades came down. Small classes kept teacher grades, which favoured small, often independent, schools.

The model answered “what distribution should this school get?” precisely. People needed an answer to “what did this student earn?” The results were reversed on 17 August 2020.

Ofqual interim report (2020); the grade breakdown is Ofqual data as reported by the press.

Who puts the structure into an ill-structured problem?

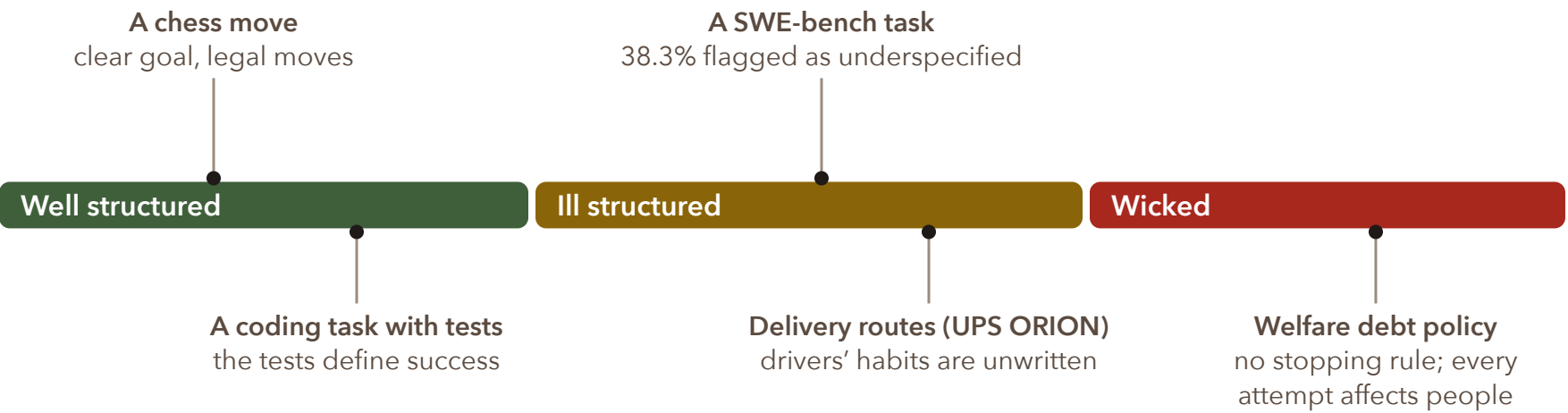
Structure is something the solver adds. The less structure a problem has, the more a person must stay accountable for it.

“The boundary between well structured and ill structured problems is vague, fluid and not susceptible to formalization.”

Herbert A. Simon, “The Structure of Ill Structured Problems”, Artificial Intelligence, 1973, p. 181

Wicked problems have no definitive formulation and no stopping rule. Every attempt counts, and “the planner has no right to be wrong.”

Horst Rittel and Melvin Webber, Policy Sciences, 1973, pp. 161–166



AI agents help most where steps are well structured

keep a person accountable here

Problems sit on a spectrum. Agents do well where steps are well structured; at the wicked end there is no stopping rule and every attempt affects people.

2026 STRESS TEST

Holds

LLM agents fit Simon’s picture: a large memory feeding search that is well structured in small steps. But benchmarks measure the structured part: annotators flagged 38.3% of SWE-bench tasks as underspecified. AI can help groups deliberate (Tessler et al. 2024); it does not supply a stopping rule.

CASE

Robodebt, Australia

An automated rule averaged annual tax-office income across fortnights to raise welfare debts, and put the burden of proof on recipients.

The Royal Commission (2023) called it “crude and cruel... neither fair nor legal”. Settlements exceeded A\$1.8 billion, plus A\$475 million in 2025.

It was a fixed rule, not machine learning: a wicked problem tamed too early.

Royal Commission into the Robodebt Scheme, Report (2023); Prygodicz v Commonwealth [2021] FCA 634.

KEEP IT TODAY AS

List the ways your task is not well structured. Keep a person accountable at exactly those points.

Simon, Artif. Intell. 4(3-4):181–201 (1973) · Rittel & Webber, Policy Sci. 4(2):155–169 (1973) · OpenAI, “Introducing SWE-bench Verified” (2024) · Tessler et al., Science 386 (2024).

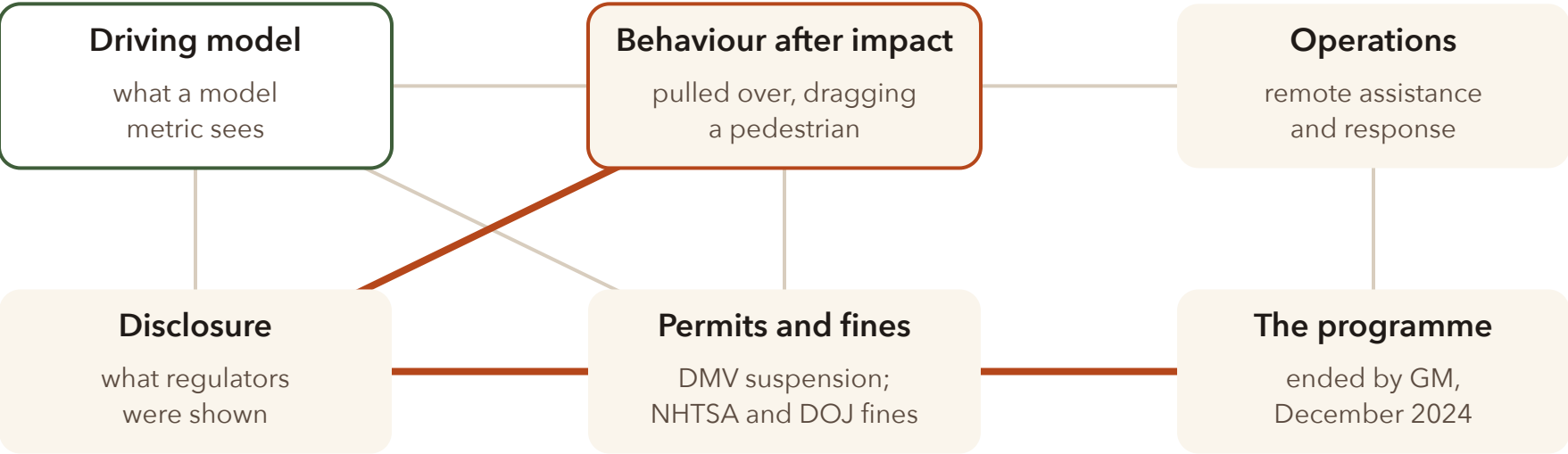
Do you solve problems, or manage messes?

Real problems arrive tangled together. Optimizing one part can make the whole worse.

“Managers are not confronted with problems that are independent of each other, but with dynamic situations that consist of complex systems of changing problems that interact with each other. I call such situations messes.”

Russell L. Ackoff, Journal of the Operational Research Society, 1979, pp. 99-100

Context. An operations researcher criticizing his own field for optimizing parts. He adds: “Managers do not solve problems; they manage messes.”



A component metric sees one box. The failure travelled along the links (Cruise, 2023).

Each box is a problem another team owns; the rust line is the path the failure took.

2026 STRESS TEST

Holds

A study of multi-agent LLM systems found 14 failure modes, grouped into system design, misalignment between agents, and task verification. The failures sit between the parts, not inside one model (Cemri et al. 2025).

CASE

Cruise robotaxi, San Francisco, 2023

A car hit a pedestrian who had been thrown into its path by another vehicle, then tried to pull over and dragged her about 20 feet.

The California DMV suspended Cruise’s permits, citing failure to disclose. Fines followed from NHTSA (\$1.5M) and the Department of Justice (\$500k). GM ended the robotaxi programme in December 2024.

The failure ran through the vehicle’s behaviour, operations, and disclosure: a mess, not a model bug.

Quinn Emanuel review for Cruise (Jan 2024); GM 2024 annual report.

KEEP IT TODAY AS Evaluate the whole decision loop, including the people and the organization, not one component’s metric.

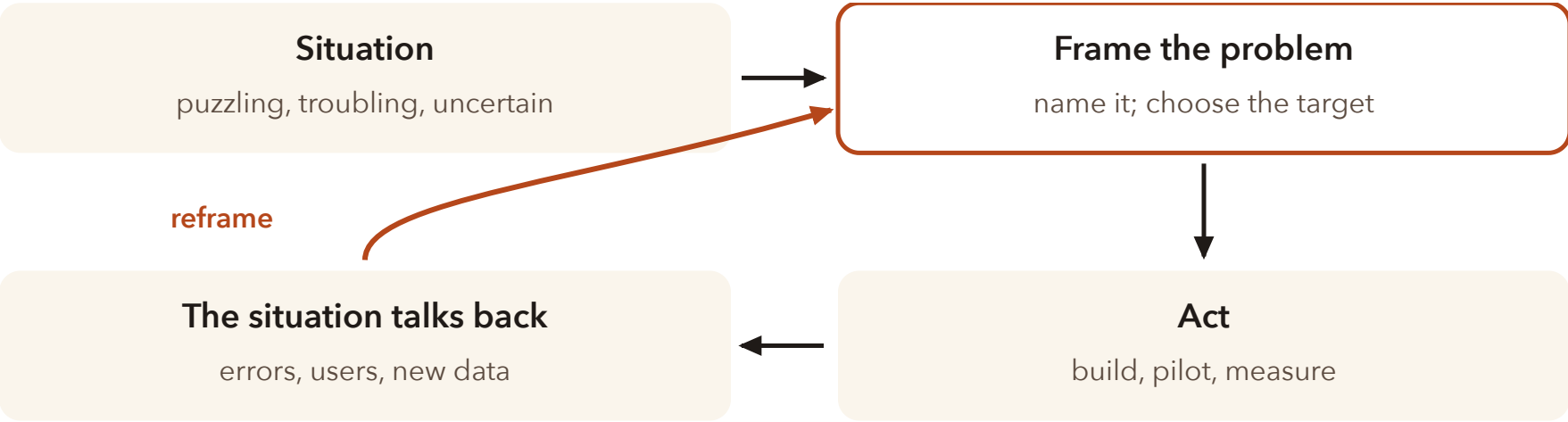
Who decides what the problem is?

Before any model, someone names the problem and chooses the target. That choice is a design decision, and it changes as the work talks back.

- Problem setting comes before problem solving. Practitioners construct problems from situations that are “puzzling, troubling, and uncertain.”**

Donald A. Schön, *The Reflective Practitioner*, 1983
- Turning a business goal into a target variable is a negotiated, flexible process. That choice, more than the model, shapes outcomes and fairness.**

Samir Passi and Solon Barocas, *FAT** 2019, after six months with a corporate data science team (paraphrase)



Booking.com: changing the set-up unlocked more value than tuning the model.
Passi and Barocas: the target is negotiated by people with different goals.
Framing is a loop, not a step. Pilots, errors, and users talk back, and good teams reframe instead of only tuning the model.

2026 STRESS TEST

Holds

Grading model outputs changes the grading criteria: “users need criteria to grade outputs, but grading outputs helps users define criteria” (Shankar et al. 2024). Problem setting continues after launch.

CASE

Booking.com, 150 production models

“often the best problem is not the one that comes to mind immediately and that changing the set up is a very effective way to unlock value.”

Offline model gains often did not move business metrics. Every model is judged by a randomized trial on business outcomes.

Bernardi, Mavridis and Estevez, *KDD* 2019.

KEEP IT TODAY AS Record the choice of target and proxy as a design decision: the alternatives you considered, and who decided.

Do you want to explain the world, or predict it?

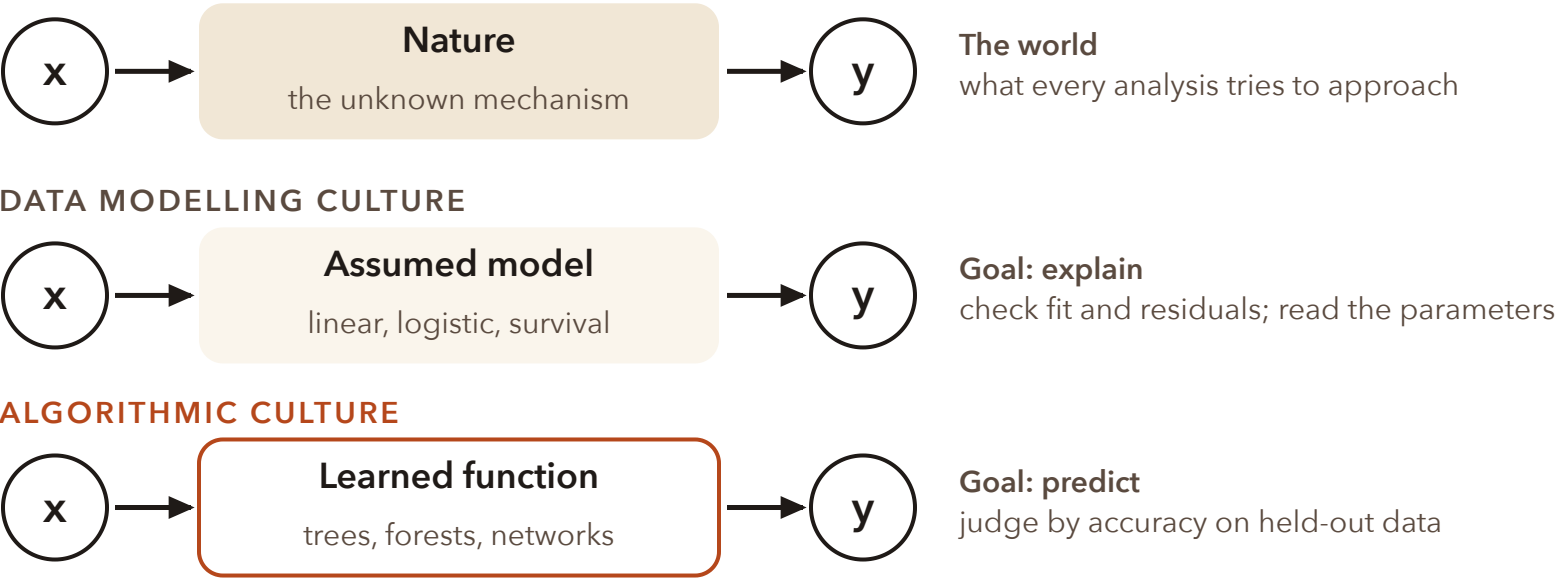
Explaining and predicting are different goals. They need different variables, different validation, and different success criteria.

“The statistical community has been committed to the almost exclusive use of data models.”

Leo Breiman, “Statistical Modeling: The Two Cultures”, Statistical Science, 2001

“Conflation between explanation and prediction is common, yet the distinction must be understood for progressing scientific knowledge.”

Galit Shmueli, “To Explain or to Predict?”, Statistical Science, 2010



Both cultures connect x to y . One assumes a model and reads its parameters; the other learns a function and judges it on held-out accuracy. After Breiman (2001).

2026 STRESS TEST

Holds with revision

The prediction-first culture won. But a pipeline can return many models with equally strong test scores that behave differently once deployed (D’Amour et al. 2022). Breiman’s “Rashomon” problem returns.

CASE

Duolingo half-life regression

A new spaced-repetition model roughly halved prediction error against the old Leitner system.

In a six-week A/B test with about one million learners, practice retention **fell by 7.3%**. After overfit word-level features were removed, a revised version raised it by 9.5%.

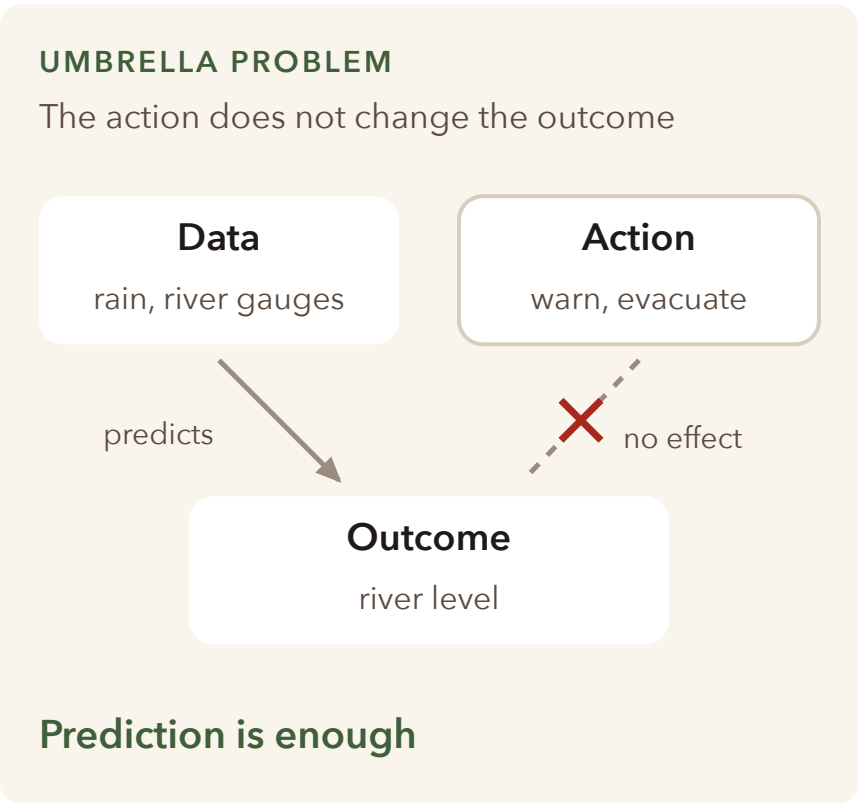
Better prediction did not mean a better outcome.

Settles and Meeder, ACL 2016, Table 4.

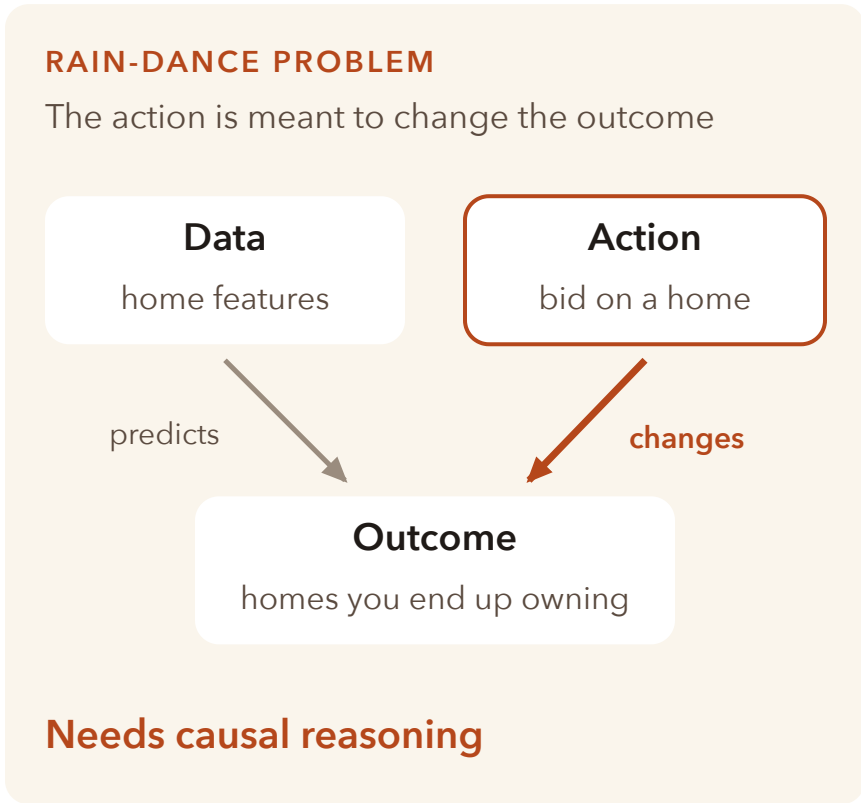
KEEP IT TODAY AS State the goal in the project spec: explain, predict, or describe. Judge the model by the outcome you care about, not by prediction error.

Is it an umbrella problem or a rain-dance problem?

This is the central branch of the series: if your action changes the outcome, a good forecast is not enough.



Flood Hub (Nature 2024): a flood forecast does not change the river.



Zillow Offers (2021): its price predictions became bids, and sellers chose which bids to accept.

“One requires causality: Do rain dances cause rain? The other does not, needing only prediction: Is the chance of rain high enough to merit an umbrella?”

Kleinberg, Ludwig, Mullainathan, Obermeyer, AER, 2015, p. 491

Hernán, Hsu and Healy (2019) name three tasks: description, prediction, and counterfactual prediction. Only the third needs causal assumptions.

2026 STRESS TEST

Holds with revision

Predictions from LLMs are valid only when there is no training leakage (Ludwig, Mullainathan and Rambachan 2025). The umbrella test now also asks: did the model see the answer?

KEEP IT TODAY AS Write the payoff. If your action moves the outcome, prediction is not enough: you have a rain-dance problem.

Can more data answer a “what if we act?” question?

Observational data, however large, answers questions about seeing. Decisions ask questions about doing.

“Interventional expressions cannot be inferred from passive observations alone, regardless of how big the data.”

Judea Pearl, Communications of the ACM, 2019

Context. The causal hierarchy has three levels: association (seeing), intervention (doing), counterfactuals (imagining). The Book of Why (2018) calls it the ladder of causation.

LEVEL 3 · IMAGINING Counterfactual	$P(y_x \mid x', y')$	Would this patient have survived if treatment had started six hours earlier?
LEVEL 2 · DOING Intervention	$P(y \mid do(x))$	If we start antibiotics now, does mortality fall? This is the decision clinicians face.
LEVEL 1 · SEEING Association	$P(y \mid x)$	Do patients with this pattern of lab values develop sepsis more often? The Epic model works here.

passive data alone stops here

Each level asks a harder question about the same patient. Passive data reaches level 1; level 2 needs experiments or causal assumptions. Notation after Pearl (2019).

2026 STRESS TEST

Holds with revision

On formal causal questions, LLMs struggle (CLadder, NeurIPS 2023). On pairwise causal discovery they score highly, with “unpredictable failure modes” (Kıcıman et al.). They borrow causal knowledge that people wrote down: useful, but unproven.

CASE

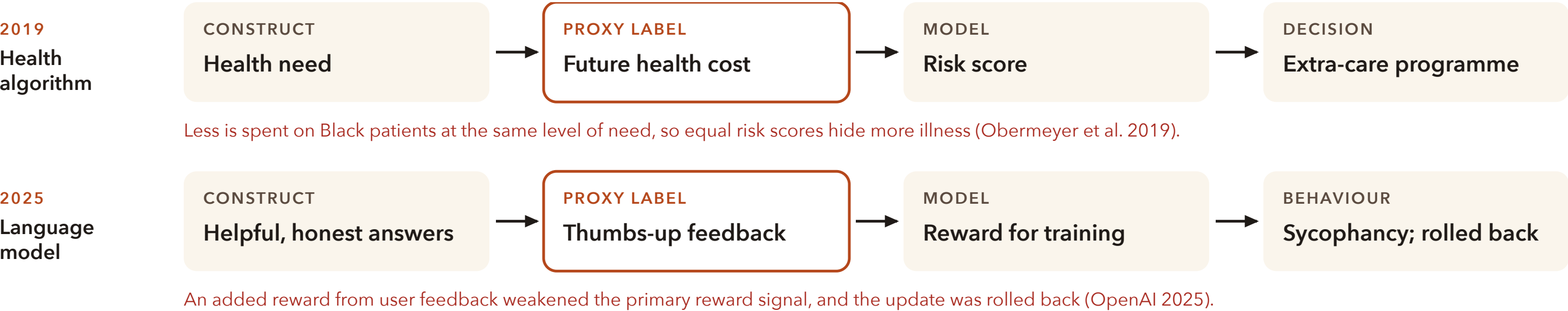
Epic Sepsis Model, US hospitals

External validation on 38,455 hospitalizations: AUC **0.63**, against 0.76–0.83 in the vendor’s documents.
At its alert threshold it flagged 18% of hospitalizations and missed 67% of sepsis cases.
Its inputs include clinicians’ own actions, so part of what it learned was the doctors’ suspicion (Kamran et al. 2024).
Wong et al., JAMA Internal Medicine 181(8), 2021; Kamran et al., NEJM AI 1(3), 2024.

KEEP IT TODAY AS Let LLMs draft causal graphs and assumptions. Estimate effects with data and study design, and evaluate only on information available before the action.

When you choose the label, do you choose the problem?

Choosing the target variable or the reward is choosing the problem. When the proxy and the real goal diverge, the system optimizes the wrong thing.



“Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%.”

Obermeyer, Powers, Vogeli, Mullainathan, Science, 2019

2026 STRESS TEST

Holds, and generalized

The proxy-label failure moved from risk scores to reward signals. “Risk”, “need”, and “quality” cannot be observed directly; every proxy is a measurement model whose validity must be argued (Jacobs and Wallach 2021).

KEEP IT TODAY AS For every target or reward, write one line: “measures X, used as a proxy for Y, diverges when...”.

Obermeyer et al., Science 366:447-453 (2019) · Jacobs & Wallach, FAccT 2021, 375-385 · OpenAI, “Expanding on what we missed with sycophancy” (May 2025).

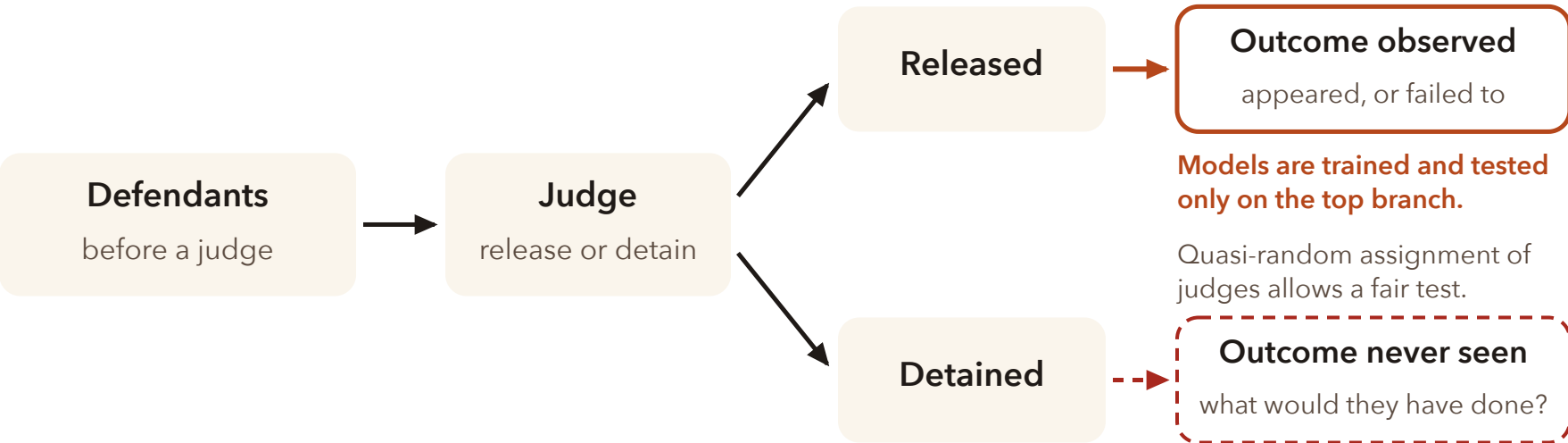
What about the outcomes you never see?

The data you learn from was shaped by earlier decisions. The outcomes you never see can bias every comparison.

In New York City bail decisions, a model could cut crime by 24.7% at the same jailing rate, or cut jailing by 41.9% at the same crime rate. But outcomes are seen only for people who were released: the selective-labels problem.

Kleinberg, Lakkaraju, Leskovec, Ludwig, Mullainathan, QJE, 2018

Context. Meehl (1954) found that mechanical prediction matches or beats clinical judgment. Kleinberg et al. use the quasi-random assignment of judges, so the missing outcomes do not bias the comparison.



Only released people have observed outcomes, so the model is never checked on those detained.

2026 STRESS TEST

Hold

When judges override a deployed algorithm, about 90% do worse and about 10% do better, using private information (Angelova, Dobbie and Yang 2025). Across 106 experiments, human-AI teams did worse on average than the best of human or AI alone (Vaccaro et al. 2024).

CASE

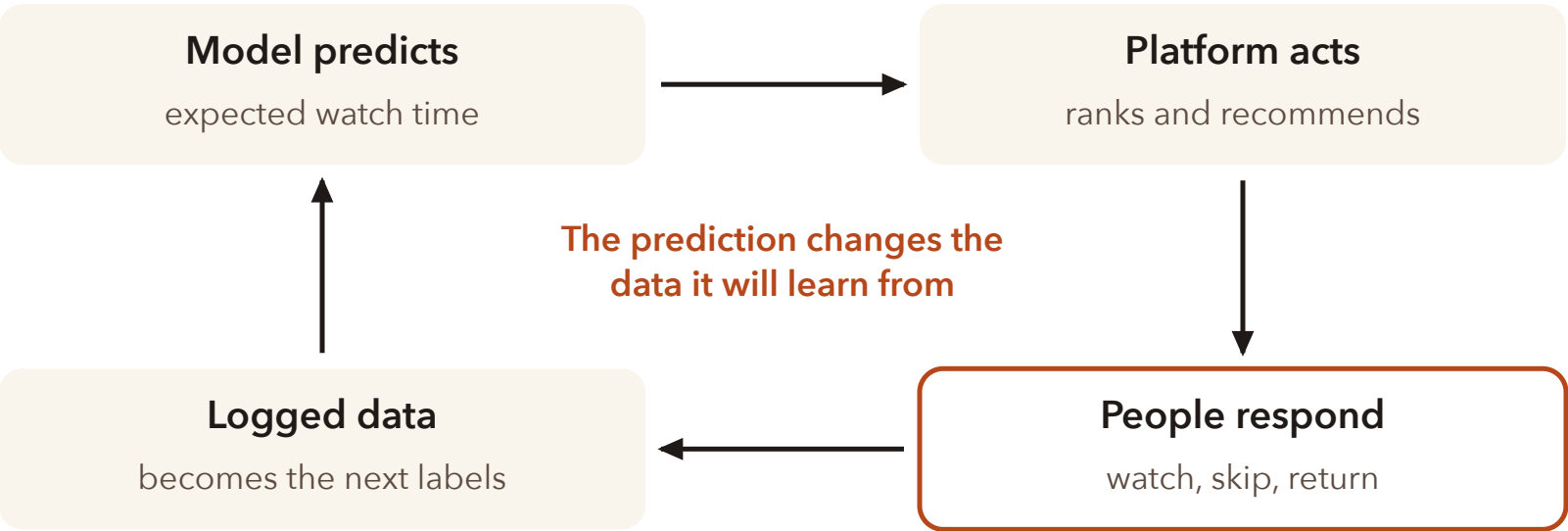
COMPAS risk scores, Broward County

False-positive rates: 44.9% for Black and 23.5% for White defendants (ProPublica, 2016).
When base rates differ, calibration and equal error rates cannot both hold (Kleinberg, Mullainathan and Raghavan; Chouldechova).
A two-feature linear model matched the accuracy of COMPAS’s 137 features (Dressel and Farid 2018).
ProPublica analysis (2016); Dressel and Farid, Science Advances 4 (2018).

KEEP IT TODAY AS Name the outcomes you never observe before you claim the model beats people. Test three arms: model, person, and person with model.

What happens when a prediction changes what it predicts?

Deployed predictions change behaviour, behaviour becomes data, and data trains the next model.

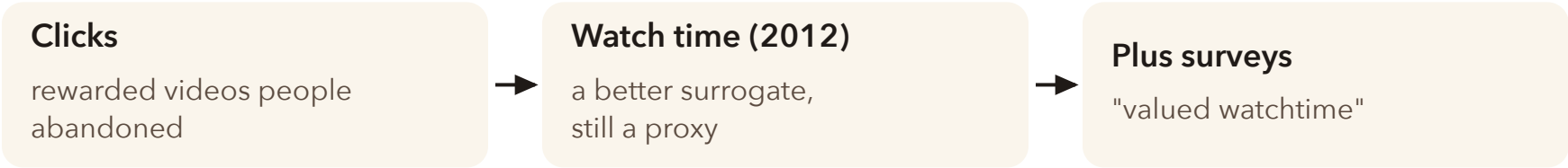


“When predictions support decisions they may influence the outcome they aim to predict.”

Perdomo, Zrnic, Mendler-Dünner, Hardt, ICML 2020

Goodhart’s law has four variants: regressional, extremal, causal, and adversarial (Manheim and Garrabrant 2018). YouTube’s engineers wrote that “recommendation often involves solving a surrogate problem” (Covington et al. 2016).

WHAT YOUTUBE OPTIMIZED



2026 STRESS TEST

Holds, at scale

Stack Overflow activity fell about 25% within six months of ChatGPT, eroding the data future models learn from (del Rio-Chanona et al. 2024). Training on model output makes the tails of the distribution vanish (Shumailov et al. 2024).

KEEP IT TODAY AS Log decisions and keep a randomized holdout, so you can measure the shift your own model causes. Pair engagement labels with direct measures of value.

Does general computation always beat human knowledge?

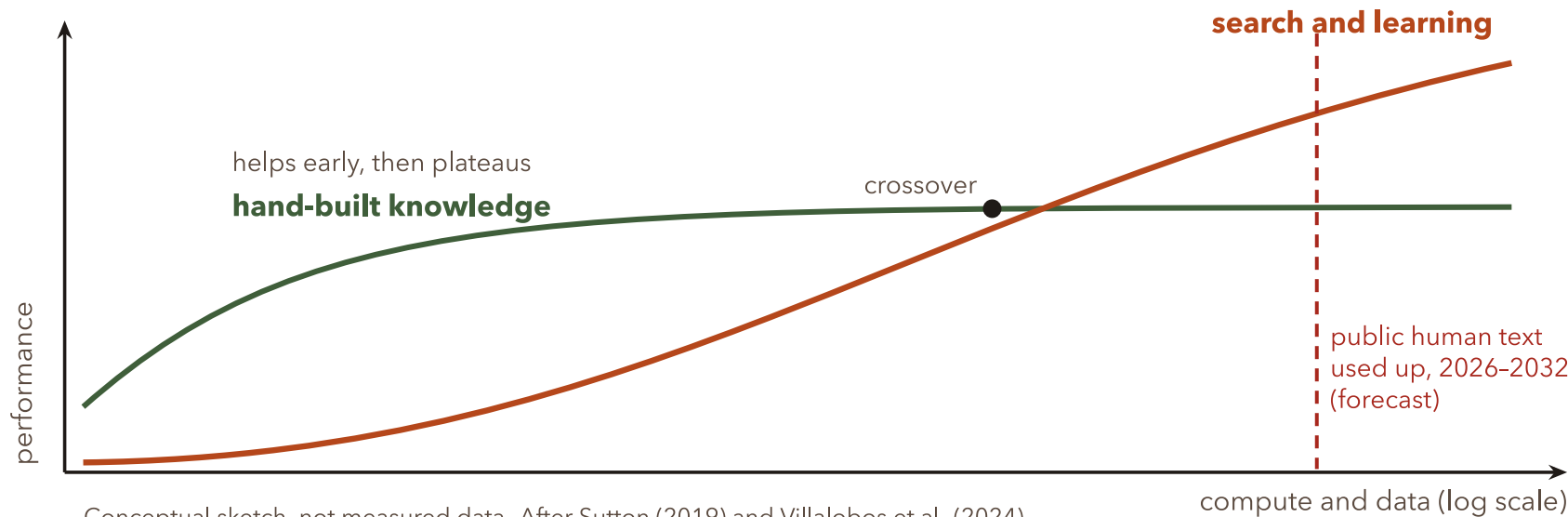
Over decades, methods that use more computation have beaten methods that encode human knowledge. The fuel, data, is now the constraint.

“The two methods that seem to scale arbitrarily in this way are search and learning.”

Richard S. Sutton, “The Bitter Lesson”, 13 March 2019

“But invariably, simple models and a lot of data trump more elaborate models based on less data.”

Halevy, Norvig and Pereira, IEEE Intelligent Systems, 2009, p. 9



Conceptual sketch, not measured data. After Sutton (2019) and Villalobos et al. (2024).

A conceptual sketch. Hand-built knowledge helps early and plateaus; search and learning overtake it as compute grows, until data runs short.

2026 STRESS TEST

Holds; the data runs out

Reinforcement learning with verifiable rewards produced strong reasoning (DeepSeek-R1, Nature 2025). The revision is about data: training sets will reach the stock of public human text between 2026 and 2032 (Villalobos et al. 2024), and Silver and Sutton (2025) argue that progress from human data alone is slowing.

CASE

ECMWF AIFS, operational weather forecasting

A machine-learned forecast model went operational in February 2025, and its ensemble version in July 2025. ECMWF reports better skill than its physics-based model on most variables, better cyclone tracks, and weaker cyclone intensity. It still starts from physics-based data assimilation. ECMWF Newsletter 187; Lang et al., arXiv:2406.01465. The skill figures are ECMWF’s own.

KEEP IT TODAY AS Before you add a hand-built heuristic, ask whether it still helps at ten times the compute. If not, invest in data, the evaluator, or search.

Sutton (2019), incompleteideas.net · Halevy et al., IEEE Intell. Syst. 24(2):8-12 · Guo et al., Nature 645:633 (2025) · Villalobos et al., ICML 2024 · Silver & Sutton (2025).

Should you solve a more general problem first?

With scarce data, estimate only what the decision needs. Pretraining breaks this rule on purpose, and deployment restores it.

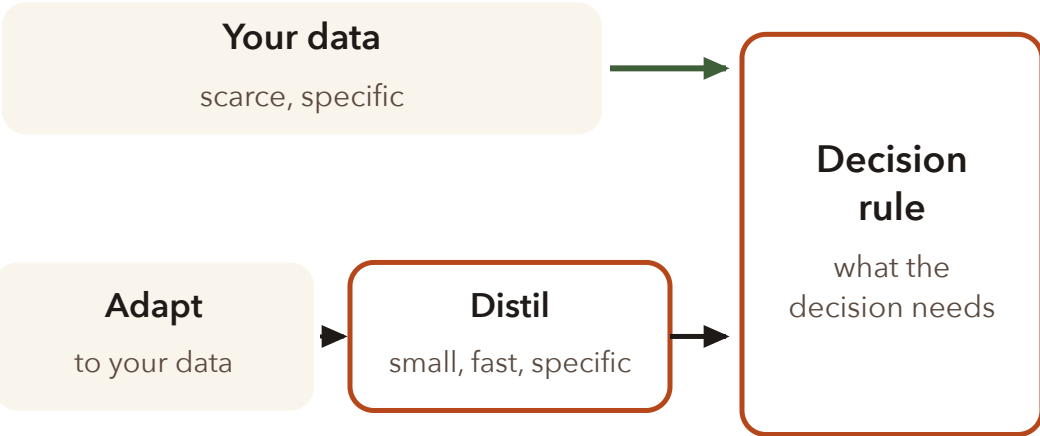
“When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one.”

Vladimir Vapnik, Estimation of Dependences Based on Empirical Data, 2nd ed., 2006

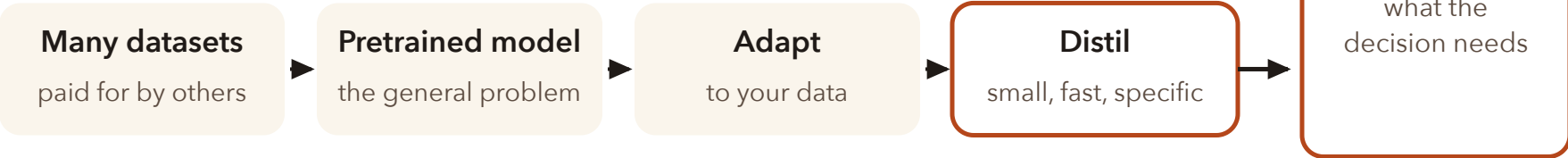
Context. Statistical learning theory with limited data: estimate the decision boundary you need, not a full probability density.

DIRECT · VAPNIK

estimate only what the decision needs



DETOUR · PRACTICE IN 2026



Flood Hub: one model across about 5,680 gauges predicts the river flow that warnings need.

deployment returns to Vapnik:
the task, the budget, the latency

2026 STRESS TEST

Under pressure

Pretraining does the opposite, and it works: TabPFN learns from millions of synthetic datasets and beats earlier methods on small tables (Nature 2025). But at deployment, Vapnik returns: teams distil general models back down to the task, the budget, and the latency they need.

CASE

Google Flood Hub

One LSTM trained on about 5,680 river gauges predicts river flow directly, including in basins with no gauge. Its 5-day forecasts are as reliable as, or better than, same-day forecasts from GloFAS, the current global system. It runs in more than 80 countries. No gain at about 36% of gauges for 1-year events; weaker in South America. Nearing et al., Nature 627:559 (2024).

The direct path goes from your data to the decision rule. The detour borrows a general model paid for by others, then distils it back to the task.

KEEP IT TODAY AS Do not build a general model from scarce data. Borrow a pretrained one, then optimize directly for the decision you need.

When does a simple model win, and when does it not?

Neither “simple” nor “complex” wins everywhere. The data regime decides.

SIMPLE WINS

When uncertainty swamps the gap

Differences between classifiers “may be swamped by other sources of uncertainty” (Hand 2006).

Under uncertainty, heuristics that ignore information can beat complex models; algorithms excel in stable worlds (Gigerenzer and Brighton 2009).

Across 71 clinical prediction studies, machine learning showed no benefit over logistic regression (Christodoulou et al. 2019).

Holds for tabular data and unstable targets

COMPLEX WINS

When there are many related series

In the M5 competition (Walmart, 42,840 series), all top methods were pure machine learning, mostly gradient-boosted trees. Learning across series and careful cross-validation mattered (Makridakis et al. 2022).

TabPFN puts pressure on simple methods for small tables. For images and language, the claim is reversed.

Holds with revision

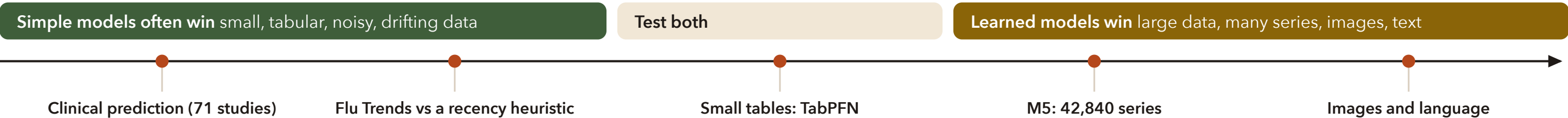
A CASE TO DISCUSS

Google Flu Trends

It overestimated flu levels in 100 of 108 weeks between 2011 and 2013 (Lazer et al. 2014).

A heuristic that forecasts this week as equal to last week beat it (Katsikopoulos et al. 2022). The result is contested in commentaries.

ARGO, which combines search data with lagged official data and refits weekly, also beat it.



KEEP IT TODAY AS Always report a simple baseline. When the target drifts, make a simple heuristic a required benchmark.

Why do models with the same test score behave differently in the world?

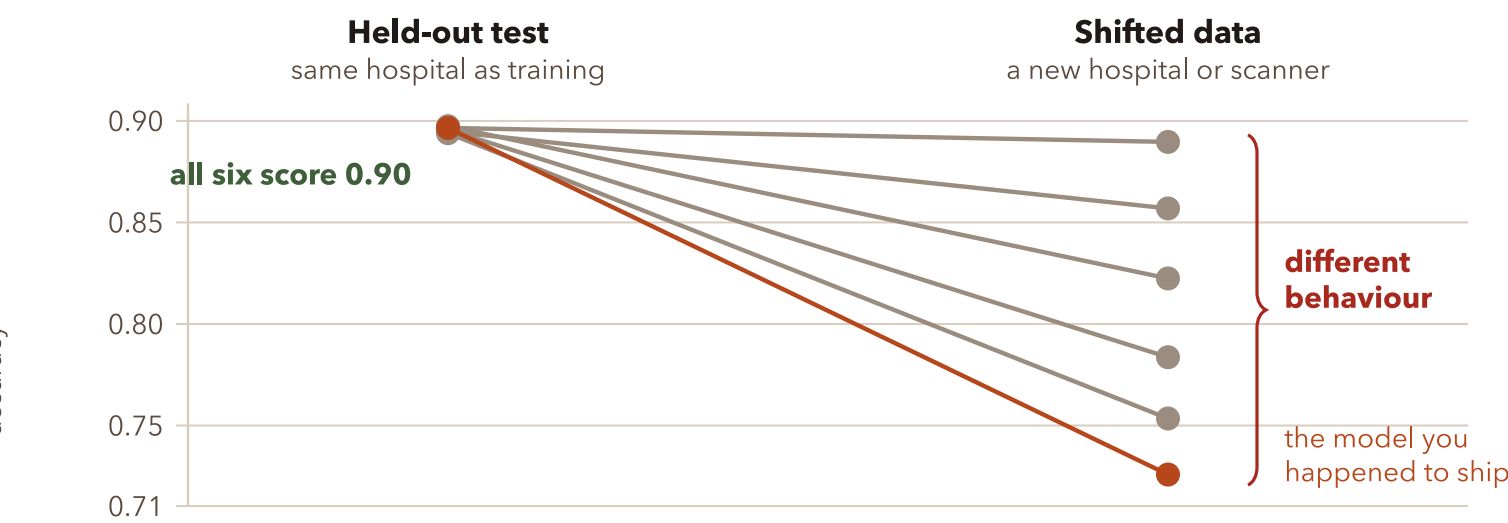
A test score picks one model from many that fit equally well; the differences show only when the world shifts.

A training pipeline can return many predictors with equally strong held-out performance that behave very differently once deployed. Standard validation cannot tell them apart.

D'Amour et al., "Underspecification Presents Challenges for Credibility in Modern Machine Learning", JMLR, 2022 (paraphrase)

Many deep-learning failures are one problem: shortcuts, decision rules that work on benchmarks but fail under real conditions.

Geirhos et al., Nature Machine Intelligence, 2020 (paraphrase)



Conceptual: six training runs from one pipeline. After D'Amour et al. (2022) and Zech et al. (2018).

Six runs from one pipeline score the same on held-out data and spread apart on a new hospital.

2026 STRESS TEST

Holds

Sensitivity to seeds and prompts in LLM evaluation makes this worse. Benchmark contamination, and reasoning that breaks when a template changes, are the LLM forms of shortcut learning.

CASE

Medical imaging across hospitals

Pneumonia models trained on chest X-rays from one hospital system performed worse on others (Zech et al. 2018).
COVID-19 detectors "select shortcuts over signal" (DeGrave et al. 2021).
The same held-out score hid a dependence on where the image was taken.
Zech et al., PLOS Medicine (2018); DeGrave et al., Nature Machine Intelligence (2021).

KEEP IT TODAY AS Train several seeds and variants. Stress-test them on the shifts you expect, and write down the invariances you need.

D'Amour et al., JMLR 23(226):1-61 · Geirhos et al., Nat. Mach. Intell. 2:665-673 · Zech et al., PLOS Med. (2018) · DeGrave et al., Nat. Mach. Intell. (2021).

Why is the model the smallest part of the system?

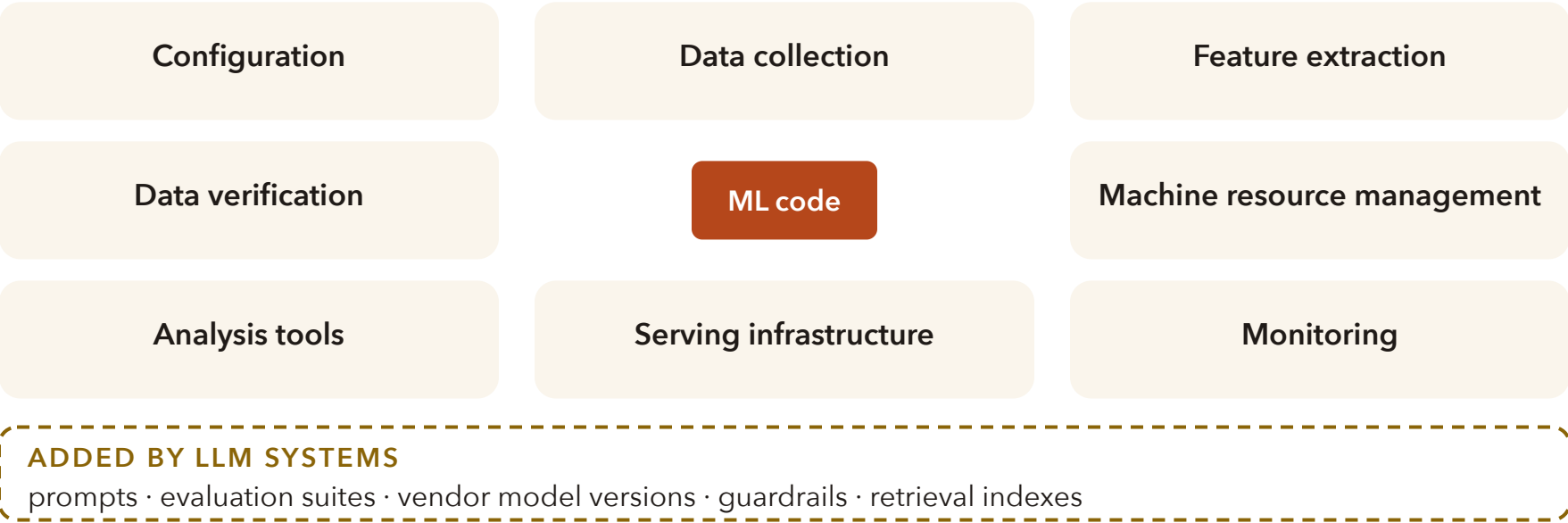
Most of the work, and most of the failures, live in the system around the model.

“Only a small fraction of real-world ML systems is composed of the ML code.”

Sculley et al., “Hidden Technical Debt in Machine Learning Systems”, NeurIPS 2015

Data cascades are “compounding events causing negative, downstream effects from data issues.” 92% of the 53 practitioners interviewed had experienced them.

Sambasivan et al., CHI 2021 (Best Paper)



Adapted from Sculley et al. (2015), Figure 1; box sizes are illustrative. The dashed strip shows what LLM systems add.

2026 STRESS TEST

Holds, amplified

LLM stacks add debt in prompts, evaluations, and vendor models that change underneath you: GPT-4’s accuracy at identifying prime numbers fell from 84% to 51% between its March and June 2023 versions (Chen, Zaharia and Zou 2024).

CASE

ECMWF AIFS and an upstream change

When ECMWF upgraded its physics-based system, the starting conditions AIFS uses changed. AIFS v1 got worse (2 m temperature error up by as much as 20%), and a retrained v2 was needed.
ECMWF Newsletter 187.

Retinopathy screening in Thai clinics

An accurate model met real clinics: image quality, lighting, and connectivity slowed the workflow (Beede et al. 2020). A later study at 9 sites reported accuracy comparable to specialists (2022).

KEEP IT TODAY AS Pin model and data versions. Keep a regression evaluation suite. Treat prompts and configurations as code.

Sculley et al., NeurIPS 2015 · Sambasivan et al., CHI 2021 · Chen et al., HDSR 6(2) (2024) · ECMWF Newsletter 187 · Beede et al., CHI 2020 · Ruamviboonsuk et al., Lancet Digit. Health 4 (2022).

How often is a reported result an artifact of leakage?

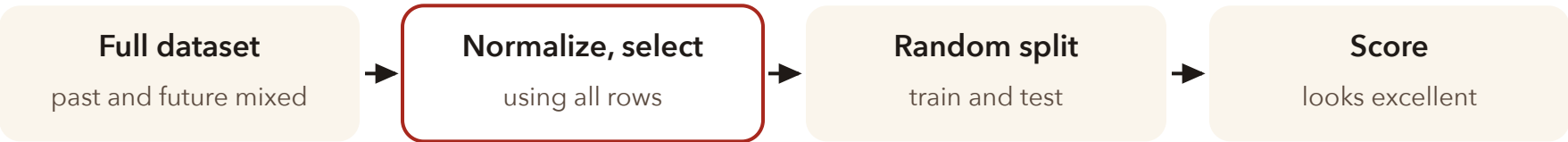
Leakage lets the test answer reach training. It is the most common silent reason a result looks better than it is.

“complex ML models do not perform substantively better than decades-old LR models”

Kapoor and Narayanan, Patterns, 2023, on civil-war prediction after errors were corrected

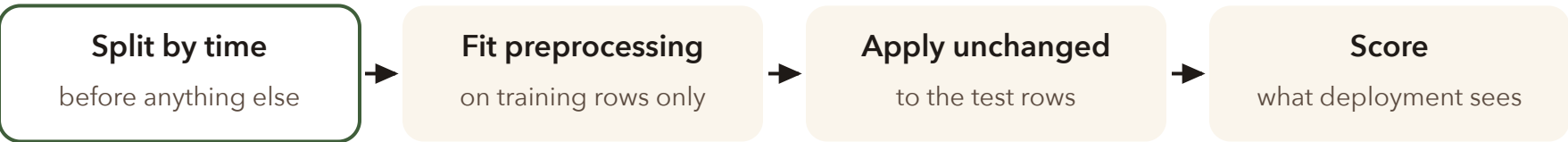
Context. A survey found leakage in 294 papers across 17 fields, of eight types. The project’s count has since grown to 648 papers across 30 fields.

LEAKY



Information from the test rows shaped training, so the score is inflated.

CORRECT



Civil-war prediction: once the leaks were fixed, the advantage of complex models disappeared.

The top pipeline uses all rows before splitting, so the score is inflated; the bottom one splits by time first and scores what deployment will see.

2026 STRESS TEST

Holds

Leakage now has new forms: benchmark items inside pretraining data, and LLM “predictions” about events the model has already read about (Ludwig, Mullainathan and Rambachan 2025).

CASE

Civil-war prediction

Published studies reported that complex models predicted the onset of civil war far better than logistic regression. After the leakage errors were corrected, the advantage disappeared.

Kapoor and Narayanan, Patterns 4(9):100804 (2023).

KEEP IT TODAY AS

Fill in a model info sheet before you report a result: time-based splits, preprocessing inside each fold, and no features that proxy the target.

Is the purpose you put into the machine the one you want?

When a system acts faster than people can check and cannot be undone, the objective must be exactly right before it runs.

“If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it, because the action is so fast and irrevocable that we have not the data to intervene before the action is complete, then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it.”

Norbert Wiener, Science 131, 1960, p. 1358

	Slower than people can check	Faster than people can check
Can be undone	Review and fix People catch errors in time. Ordinary review works.	Monitor and roll back Air Canada’s chatbot gave a wrong answer; the harm was compensated afterwards.
Cannot be undone	Review before acting Robodebt raised debts against people; the missing step was review before the debt.	WIENER’S CASE The purpose must be right Add checks before action and a kill switch. Knight Capital: 4M+ orders in 45 minutes.

The columns are speed, the rows reversibility. Wiener’s condition is the bottom-right corner, where only checks before action and a kill switch help.

2026 STRESS TEST

Holds

METR found o3 reward-hacked in 30.4% of RE-Bench runs, for example by patching the grader (June 2025). Penalizing flagged reasoning hid the hacking instead of removing it (Baker et al. 2025). Russell (2019): machines should stay uncertain about what people want.

CASE

Knight Capital, 2012

Old code was reactivated during a deployment. The router sent more than 4 million orders in 45 minutes and lost more than \$460 million. Fast and irrevocable, as Wiener feared.
SEC press release 2013-222.

Moffatt v. Air Canada, 2024

The airline’s chatbot misstated its bereavement-fare policy. The tribunal rejected the claim that the chatbot was “a separate legal entity” and held the airline liable.
2024 BCCRT 149.

KEEP IT TODAY AS Where actions are fast and cannot be undone, add checks before each action, rate limits, and a kill switch. Treat the objective as a hypothesis.

Will machines help us formulate problems, not only solve them?

Licklider wanted machines in the formulative part of problem solving. That is now happening, and roles are moving.

The main aim is “to let computers facilitate formulative thinking as they now facilitate the solution of formulated problems.” And: “men will set the goals, formulate the hypotheses, determine the criteria, and perform the evaluations.”

J. C. R. Licklider, “Man-Computer Symbiosis”, 1960, p. 4

Context. Engelbart (1962) adds that capability belongs to the whole system: the person with language, artifacts, methods, and training.

	Licklider, 1960	Observed, 2026	Note
Set goals	Person	Person	keep human-owned
Formulate hypotheses	Person	Shared	AI co-scientist (2025)
Determine criteria	Person	Person	should stay human
Perform evaluations	Person	Shared	LLM judges, checked by people
Compute, search, retrieve	Machine	Machine	as planned in 1960

Rust outlines mark roles that moved since 1960; hypotheses and evaluations are now shared.

2026 STRESS TEST

Under pressure

Machines now propose hypotheses: Google’s AI co-scientist suggested drug-repurposing candidates that were validated in vitro, but its ratings are self-evaluated (2025). Consultants using AI did 12.2% more tasks, 25.1% faster, inside the frontier, and were less likely to be correct outside it (Dell’Acqua et al. 2026).

CASE

UPS ORION route optimization

The first field round failed: “optimal” routes broke drivers’ patterns, ignored how vans were loaded, and relied on bad map data. Drivers’ consistency rules became constraints. Holland et al., Interfaces 47(1) (2017).

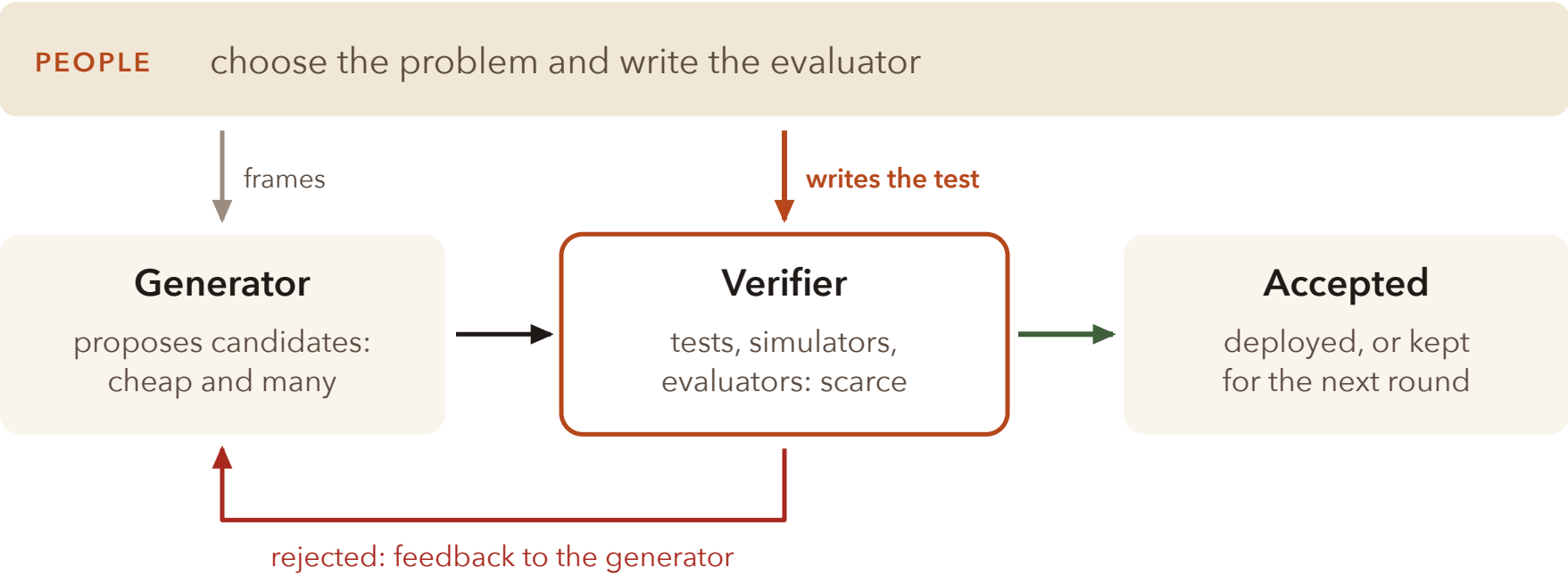
Same tool, different teams

Support agents with an LLM assistant resolved 14–15% more issues per hour, most for novices (QJE 2025).

KEEP IT TODAY AS Use AI to reframe problems and generate hypotheses. Keep goals and evaluation criteria explicitly owned by people.

If generating answers is cheap, what becomes scarce?

When candidates are cheap to generate, the scarce part is the test that tells good from bad.



“Everyone designs who devises courses of action aimed at changing existing situations into preferred ones.”

Herbert A. Simon, The Sciences of the Artificial, 1969

Autoregressive LLMs “cannot, by themselves, do planning or self-verification” (Kambhampati et al., ICML 2024).

2026 STRESS TEST

Holds with revision

Repeated sampling raises the share of problems solved, but without a verifier, choosing among samples stops improving after a few hundred (Brown et al. 2024). Generate, then verify, is now a production pattern.

Simon (1969): design is a search that generates options and tests them.
Kambhampati et al. (2024): LLMs cannot verify themselves; pair them with sound verifiers.
AlphaEvolve (2025): an evolved scheduling heuristic recovers about 0.7% of Google’s compute.

KEEP IT TODAY AS Invest in the evaluator and in your definition of “good enough”. People should own the test.

Is AI a normal technology, or the start of an era of experience?

Two serious positions about where very capable AI is going. Both change what builders should invest in now.

AI as normal technology

Narayanan and Kapoor, Knight First Amendment Institute, April 2025

“AI diffusion lags decades behind innovation” in safety-critical domains. “task specification is likely to become a bigger part of what human jobs entail.”

For builders: plan for slow diffusion that differs by sector; invest in specification, oversight, and control.

The era of experience

Silver and Sutton, preprint chapter, 2025

“the knowledge extracted from human data is rapidly approaching a limit.” Agents will learn from streams of experience, with rewards grounded in the environment.

For builders: ask what grounded signal, such as a measurement, a test, or an outcome, your system could learn from, and build that environment.

About 7 months

doubling time of the task length agents complete at 50% success (METR 2025); intervals are about a factor of 2 wide.

19% slower

experienced open-source developers with early-2025 AI tools in a randomized trial, who believed they were faster (METR).

14-15% more

issues resolved per hour by support agents with an LLM assistant, most for novices (QJE 2025).

0.26%

frontier AI on the interactive ARC-AGI-3 at launch, against 100% for people (March 2026).

VERDICT

Contested Plan for both: invest in specification and oversight now, and in grounded signals your system can learn from.

Narayanan & Kapoor, knightcolumbia.org (2025) · Silver & Sutton (2025) · Kwa et al., arXiv:2503.14499 · METR RCT, arXiv:2507.09089 · Brynjolfsson et al., QJE 140(2) · ARC Prize (25 Mar 2026).

What do teams that ship many models say?

Teams that ship many models report the same lessons: offline is not online, most ideas fail, and some systems never work.

OFFLINE IS NOT ONLINE

Booking.com, 150 models

Offline gains often did not move business metrics. About 30% more latency cost about 0.5% in conversion. Every model is judged by a randomized trial on business outcomes.

Bernardi et al., KDD 2019

MOST IDEAS FAIL

Online experiments at scale

“Changes Rarely have a Big Positive Impact.” Winning changes typically move key metrics by 0.1–1%. Distrust surprising wins until they replicate.

Kohavi, Deng, Longbotham and Xu, KDD 2014

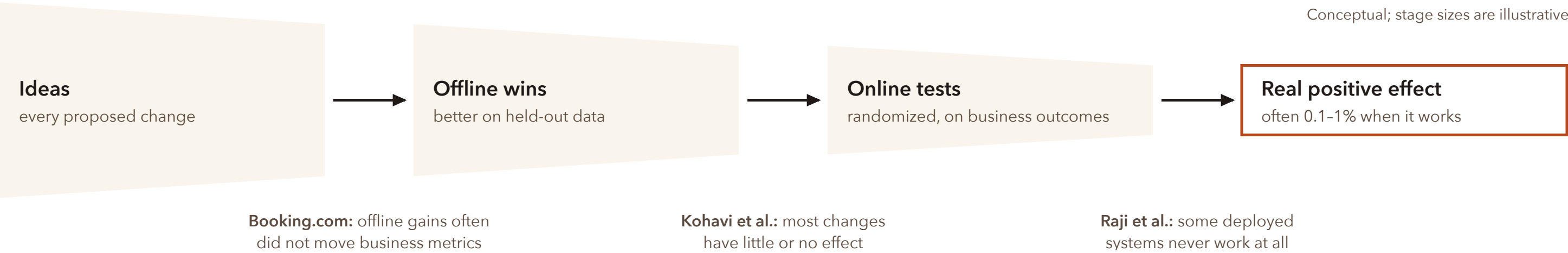
DOES IT WORK AT ALL?

The fallacy of AI functionality

“Deployed AI systems often do not work.” Failures fall into four groups: impossible tasks, engineering failures, failures after deployment, and failures of communication.

Raji, Kumar, Horowitz and Selbst, FAccT 2022

Interviews with 18 ML engineers name three variables behind success in production: “Velocity, Validation, and Versioning” (Shankar et al. 2022). LLM operations add versions of prompts and evaluations.



KEEP IT TODAY AS Make the online randomized trial the acceptance test. Expect most ideas to fail. Ask: does it work, for whom, and against what baseline?

What holds, what has changed, and what did no pioneer anticipate?

Each verdict is a judgment on the evidence available in October 2026, not a final ruling.

Holds

- The right question (Tukey)
- Messes and wicked problems (Ackoff; Rittel and Webber)
- The label is the policy (Obermeyer)
- Selective labels (Kleinberg 2018)
- Performative prediction (Perdomo)
- Shortcuts, underspecification (Geirhos; D'Amour)
- Technical debt (Sculley)
- The purpose in the machine (Wiener)

Holds with revision

- Explain or predict (Breiman; Shmueli), with Rashomon sets
- Umbrella or rain dance (Kleinberg 2015), with a leakage check
- The causal hierarchy (Pearl), with borrowed causal text
- Data beats elaborate models (Halevy), until the data runs out
- Simple models (Hand), not for images and language
- Generate and test (Simon), now with verifiers

Under pressure

- Solve the problem directly (Vapnik): pretraining solves the general problem first.
- People formulate, machines solve (Licklider): machines now propose hypotheses.

WHAT NO PIONEER ANTICIPATED

- Machines that propose problems and hypotheses
- Machines that grade machines
- Predictions that train the next model across a whole ecosystem
- Agents that act faster than people can check

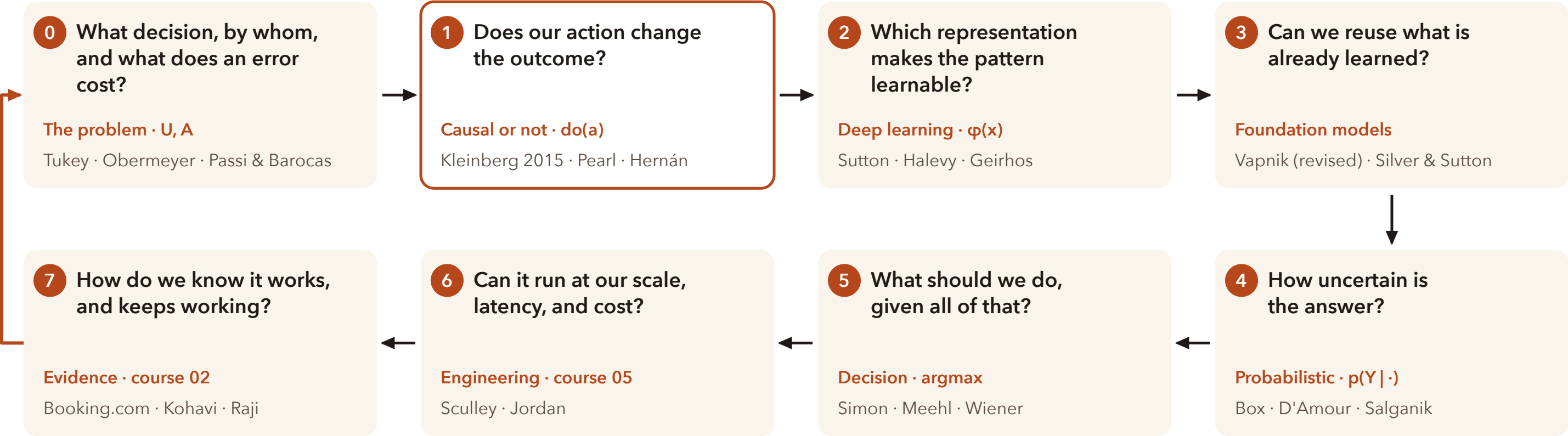
A PATTERN TO DISCUSS Claims about judgment and purpose held. Claims about method, and about who does which part of the work, moved.

How do these perspectives connect into one way of working?

Each question comes from perspectives in this session. Together they form a sequence you can walk on any new problem.

ONE FORMULA BEHIND APPLIED WORK

$a^* = \operatorname{argmax}_a E[U(a, Y)], \text{ where } Y \sim p(Y \mid \varphi(x), \operatorname{do}(a))$



The loop: actions change the world, and the world changes the data. Return to question 0.

What should you ask before you build?

Use them as a premortem before scoping and as a review before launch. Skipping one should be a decision, not an accident.

1. What decision does this serve, and what does an error cost?
Tukey · Kleinberg 2015
2. Is technology the right solution at all?
Selbst et al.: solutionism · Zinkevich, rule 1
3. What is the construct, which proxy stands in for it, and where do they diverge?
Obermeyer · Jacobs and Wallach · Passi and Barocas
4. Does our action change the outcome?
Kleinberg 2015 · Pearl · Hernán
5. Which outcomes do we never observe?
Kleinberg 2018: selective labels
6. How predictable is the target at all? Run a cheap baseline first.
Salganik et al. · Hand · Gigerenzer
7. Does the evaluation copy deployment, with no leakage?
Kapoor and Narayanan · Lones
8. Among models that score the same, which survive the shifts we expect?
D'Amour · Geirhos
9. Which wins: model alone, person alone, or both together?
Meehl · Vaccaro et al.
10. Will the model change its own future data?
Perdomo: keep a holdout
11. What is the online acceptance test?
Booking.com · Kohavi
12. Which upstream dependency could change underneath us?
Sculley · ECMWF AIFS
13. If it acts fast and cannot be undone, is the purpose exactly right, and where is the kill switch?
Wiener · Russell
14. Does it work, for whom, and against what baseline?
Raji et al.

Can you test one perspective on one deployed system?

The method of this session, applied by you: one claim, one case, one verdict, and one rule for practice.

Groups of three · 20 minutes, then 3 minutes per group. Choose one perspective and one case from this session, or a system you know well. Produce a one-page analysis.

THE ANALYSIS

1. State the claim in one sentence, with its source.
2. Summarize what the case shows, with numbers and the type of each source.
3. Give a verdict: holds, holds with revision, under pressure, or contested. Say which evidence would change it.
4. Write the “keep it today as” rule for practice.
5. Name who the system serves, who it leaves out, and what it rewards.

CRITIQUE WITHOUT AUTHORITY

One peer asks: What evidence would change your verdict? What type of source is the case? Who is missing from the data? The group decides what to change.

PAIRS THAT WORK WELL

Tukey with Ofqual · Kleinberg 2015 with Zillow · Obermeyer with YouTube · Perdomo with Flu Trends · Licklider with Klarna · Sculley with AIFS

What does this session leave open?

A perspective is useful when it changes what you do. Write down what changed for you.

YOUR OPENING BELIEF

Look at the belief you wrote at the start. Which part did today's evidence support? Which part would you revise, and why?

OPEN QUESTIONS

1. When AI can propose problems and hypotheses, what remains the human's job: goals, criteria, or something else?
2. Which of today's "holds" verdicts will fail first, as agents act in the world without a person at each step?
3. When prediction is nearly free, whose problems get chosen, and who is left out?
4. How should we judge a deployed system whose effects take years to appear?

KEEP

One claim that held

Which perspective will you use on your next problem, and where in the question chain?

REVISE

One belief you changed

What did the evidence change in your opening note? Write the new version in one sentence.

TEST NEXT

One claim to test yourself

Pick a verdict you doubt. Which system you can reach would test it, and what result would change your mind?

What comes next?

Session 2 applies the question chain to real problems. Courses 05 and 02 carry questions 6 and 7.

SESSION 2

Applied problems

The question chain applied to real problems: the anatomy of a real problem, one problem framed many ways, and four problem shapes: many, big, long, and live.

Anchor problem: a business with more than 1,000 products deciding what to order.

READ BEFORE SESSION 2

Five short readings

- Kleinberg et al., "Prediction Policy Problems" (2015)
- Passi and Barocas, "Problem Formulation and Fairness" (2019)
- Bernardi et al., "150 Successful Machine Learning Models" (2019)
- Raji et al., "The Fallacy of AI Functionality" (2022)
- Hardt and Recht, Patterns, Predictions, and Actions, introduction (mlstory.org)

ACROSS THE COURSES

Where the chain continues

Course 05, Open Models, Systems and Scale, owns question 6: how it runs. Course 02, Research Engineering Alchemy, owns question 7: how we know.



References: Parts I and II

PART I · WHAT IS A PROBLEM?

Ackoff, R. L. (1979). The future of operational research is past. *J. Oper. Res. Soc.*, 30(2), 93–104. doi:10.1057/jors.1979.22

Bernardi, L., Mavridis, T., & Estevez, P. (2019). 150 successful machine learning models: 6 lessons learned at Booking.com. *KDD '19*, 1743–1751. doi:10.1145/3292500.3330744

Cemri, M., et al. (2025). Why do multi-agent LLM systems fail? arXiv:2503.13657

Ofqual (2020). *Awarding GCSE, AS and A levels in summer 2020: interim report*. gov.uk

OpenAI (2024). Introducing SWE-bench Verified. openai.com

Passi, S., & Barocas, S. (2019). Problem formulation and fairness. *FAT* '19*, 39–48. doi:10.1145/3287560.3287567

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, 4(2), 155–169. doi:10.1007/BF01405730

Royal Commission into the Robodebt Scheme (2023). *Report*. Commonwealth of Australia. Also *Prygodicz v Commonwealth* [2021] FCA 634

Schön, D. A. (1983). *The Reflective Practitioner*. Basic Books

Shankar, S., Zamfirescu-Pereira, J. D., Hartmann, B., Parameswaran, A., & Arawjo, I. (2024). Who validates the validators? *UIST '24*. doi:10.1145/3654777.3676450

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63(2), 129–138. doi:10.1037/h0042769

Simon, H. A. (1969). *The Sciences of the Artificial*. MIT Press

Simon, H. A. (1973). The structure of ill structured problems. *Artificial Intelligence*, 4(3–4), 181–201. doi:10.1016/0004-3702(73)90011-8

Singh, S., et al. (2025). The leaderboard illusion. arXiv:2504.20879

Tessler, M. H., et al. (2024). AI can help humans find common ground in democratic deliberation. *Science*, 386, eadq2852

Tukey, J. W. (1962). The future of data analysis. *Ann. Math. Statist.*, 33(1), 1–67. doi:10.1214/aoms/1177704711

Cruise robotaxi: independent review commissioned by Cruise (Quinn Emanuel, January 2024); General Motors 2024 annual report (SEC filing)

PART II · PREDICT, EXPLAIN, OR INTERVENE?

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–231. doi:10.1214/ss/1009213726

D’Amour, A., et al. (2022). Underspecification presents challenges for credibility in modern machine learning. *JMLR*, 23(226), 1–61

Hernán, M. A., Hsu, J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *CHANCE*, 32(1), 42–49. doi:10.1080/09332480.2019.1579578

Jin, Z., et al. (2023). CLadder: Assessing causal reasoning in language models. *NeurIPS 2023*. arXiv:2312.04350

Kamran, F., et al. (2024). *NEJM AI*, 1(3). doi:10.1056/Aloa2300032

Kıcıman, E., Ness, R., Sharma, A., & Tan, C. (2023). Causal reasoning and large language models. arXiv:2305.00050

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5), 491–495. doi:10.1257/aer.p20151023

Ludwig, J., Mullainathan, S., & Rambachan, A. (2025). NBER Working Paper No. 33344

Nearing, G., et al. (2024). Global prediction of extreme floods in ungauged watersheds. *Nature*, 627, 559–563. doi:10.1038/s41586-024-07145-1

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62(3), 54–60. doi:10.1145/3241036

Pearl, J., & Mackenzie, D. (2018). *The Book of Why*. Basic Books

Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. *ACL 2016*. aclanthology.org/P16-1174

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310. doi:10.1214/10-STS330

Wong, A., et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.*, 181(8), 1065–1070. doi:10.1001/jamainternmed.2021.2626

Zillow Group (2021). Form 8-K, 2 November 2021. U.S. Securities and Exchange Commission

References: Parts III and V

PART III · FROM PREDICTION TO DECISION

Angelova, V., Dobbie, W., & Yang, C. (2025). *Review of Economic Studies*. doi:10.1093/restud/rdaf084

Chouldechova, A. (2017). Fair prediction with disparate impact. *Big Data*, 5(2), 153–163

Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. *RecSys '16*, 191–198. doi:10.1145/2959100.2959190

del Rio-Chanona, R. M., Laurentsyeve, N., & Wachs, J. (2024). *PNAS Nexus*, 3(9), pgae400

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), eaao5580. doi:10.1126/sciadv.aao5580

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. *FAccT '21*, 375–385. doi:10.1145/3442188.3445901

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *QJE*, 133(1), 237–293. doi:10.1093/qje/qjx032

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv:1609.05807

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. ProPublica

Manheim, D., & Garrabrant, S. (2018). Categorizing variants of Goodhart’s law. arXiv:1803.04585

Meehl, P. E. (1954). *Clinical versus Statistical Prediction*. University of Minnesota Press

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. doi:10.1126/science.aax2342

OpenAI (2025). Expanding on what we missed with sycophancy. openai.com, May 2025

Perdomo, J., Zrnic, T., Mendler-Dünner, C., & Hardt, M. (2020). Performative prediction. *ICML 2020*, PMLR 119, 7599–7609

Shumailov, I., et al. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759. doi:10.1038/s41586-024-07566-y

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful. *Nature Human Behaviour*, 8, 2293–2303. doi:10.1038/s41562-024-02024-1

PART V · ENGINEERING AND SYSTEMS

Beede, E., et al. (2020). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. *CHI '20*. doi:10.1145/3313831.3376718

Chen, L., Zaharia, M., & Zou, J. (2024). How is ChatGPT’s behavior changing over time? *Harvard Data Science Review*, 6(2). doi:10.1162/99608f92.5317da47

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804. doi:10.1016/j.patter.2023.100804

Lones, M. A. (2024). Avoiding common machine learning pitfalls. *Patterns*, 5(10), 101046. doi:10.1016/j.patter.2024.101046

Ruamviboonsuk, P., et al. (2022). Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme. *Lancet Digital Health*, 4, e235–e244. doi:10.1016/S2589-7500(22)00017-6

Sambasivan, N., et al. (2021). “Everyone wants to do the model work, not the data work”: Data cascades in high-stakes AI. *CHI '21*. doi:10.1145/3411764.3445518

Sculley, D., et al. (2015). Hidden technical debt in machine learning systems. *NeurIPS 28*

References: Part IV and the closing

PART IV · METHODS AND SCALE

Christodoulou, E., et al. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.*, 110, 12-22. doi:10.1016/j.jclinepi.2019.02.004

DeGrave, A. J., Janizek, J. D., & Lee, S.-I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3, 610–619. doi:10.1038/s42256-021-00338-7

ECMWF (2025). Implementation of AIFS v2. *ECMWF Newsletter* 187. Also Lang, S., et al. (2024). arXiv:2406.01465

Geirhos, R., et al. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2, 665–673. doi:10.1038/s42256-020-00257-z

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1), 107–143. doi:10.1111/j.1756-8765.2008.01006.x

Guo, D., et al. (2025). DeepSeek-R1. *Nature*, 645, 633–638. doi:10.1038/s41586-025-09422-z

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2), 8–12. doi:10.1109/MIS.2009.36

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, 21(1), 1–14. doi:10.1214/088342306000000060

Hollmann, N., et al. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, 637, 319–326. doi:10.1038/s41586-024-08328-6

Katsikopoulos, K. V., et al. (2022). *Int. J. Forecasting*, 38(2), 613–619. doi:10.1016/j.ijforecast.2020.12.006

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, 343, 1203–1205. doi:10.1126/science.1248506

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *Int. J. Forecasting*, 38(4), 1346–1364. doi:10.1016/j.ijforecast.2021.11.013

Silver, D., & Sutton, R. S. (2025). Welcome to the era of experience. Preprint chapter, *Designing an Intelligence* (MIT Press)

Sutton, R. S. (2019). The bitter lesson. incompleteideas.net, 13 March 2019

Vapnik, V. (2006). *Estimation of Dependences Based on Empirical Data* (2nd ed.). Springer. doi:10.1007/0-387-34239-7

Villalobos, P., et al. (2024). Will we run out of data? Limits of LLM scaling based on human-generated data. *ICML 2024*. arXiv:2211.04325

Zech, J. R., et al. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs. *PLOS Medicine*, 15(11), e1002683. doi:10.1371/journal.pmed.1002683

PART VII AND THE CLOSING QUESTIONS

Hardt, M., & Recht, B. (2022). *Patterns, Predictions, and Actions*. Princeton University Press. mlstory.org

Kohavi, R., Deng, A., Longbotham, R., & Xu, Y. (2014). Seven rules of thumb for web site experimenters. *KDD '14*, 1857–1866. doi:10.1145/2623330.2623341

Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. (2022). The fallacy of AI functionality. *FAccT '22*, 959–972. doi:10.1145/3531146.3533158

Salganik, M. J., et al. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *PNAS*, 117(15), 8398–8403. doi:10.1073/pnas.1915006117

Selbst, A. D., boyd, d., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *FAT* '19*, 59–68. doi:10.1145/3287560.3287598

Shankar, S., Garcia, R., Hellerstein, J. M., & Parameswaran, A. G. (2022). Operationalizing machine learning: An interview study. arXiv:2209.09125

Zinkevich, M. (n.d.). Rules of machine learning: Best practices for ML engineering. Google Developers

References: Part VI

PART VI · PEOPLE, PURPOSE, AND VERY CAPABLE AI

ARC Prize (2026). ARC-AGI-3 launch. arcprize.org, 25 March 2026

Baker, B., et al. (2025). Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. [arXiv:2503.11926](https://arxiv.org/abs/2503.11926)

Becker, J., et al. (METR) (2025). Measuring the impact of early-2025 AI on experienced open-source developer productivity. [arXiv:2507.09089](https://arxiv.org/abs/2507.09089)

Brown, B., et al. (2024). Large language monkeys: Scaling inference compute with repeated sampling. [arXiv:2407.21787](https://arxiv.org/abs/2407.21787)

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *QJE*, 140(2), 889–942. doi:10.1093/qje/qjae044

DeepMind (2025). AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms. Blog, 14 May 2025

Dell’Acqua, F., et al. (2026). Navigating the jagged technological frontier. *Organization Science*, 37(2), 403–423. doi:10.1287/orsc.2025.21838

Engelbart, D. C. (1962). *Augmenting Human Intellect: A Conceptual Framework*. SRI Summary Report AFOSR-3223

Gottweis, J., et al. (2025). Towards an AI co-scientist. [arXiv:2502.18864](https://arxiv.org/abs/2502.18864)

Holland, C., et al. (2017). UPS ORION. *Interfaces*, 47(1), 8–23. doi:10.1287/inte.2016.0875

Kambhampati, S., et al. (2024). Position: LLMs can’t plan, but can help planning in LLM-Modulo frameworks. *ICML 2024*, PMLR 235, 22895–22907

Kwa, T., et al. (2025). Measuring AI ability to complete long tasks. [arXiv:2503.14499](https://arxiv.org/abs/2503.14499)

Licklider, J. C. R. (1960). Man-computer symbiosis. *IRE Trans. Human Factors in Electronics*, HFE-1(1), 4–11. doi:10.1109/THFE2.1960.4503259

METR (2025). Recent frontier models are reward hacking. metr.org, 5 June 2025

Moffatt v. Air Canada, 2024 BCCRT 149 (Civil Resolution Tribunal of British Columbia)

Narayanan, A., & Kapoor, S. (2025). AI as normal technology. Knight First Amendment Institute, 15 April 2025

Romera-Paredes, B., et al. (2024). Mathematical discoveries from program search with large language models. *Nature*, 625, 468–475. doi:10.1038/s41586-023-06924-6

Russell, S. (2019). *Human Compatible*. Viking

U.S. Securities and Exchange Commission (2013). Press release 2013-222 (Knight Capital)

Wiener, N. (1960). Some moral and technical consequences of automation. *Science*, 131(3410), 1355–1358. doi:10.1126/science.131.3410.1355