



Knowledge With an Expiry Date

Evidence, Evaluation and Decision-Making When the AI System Keeps Changing

About this paper

This paper is written for directors, executives and the people who prepare evidence for them, which in most organisations means risk, audit, procurement, data and transformation leads. They are being asked to approve, fund or stop uses of AI on the strength of pilots, supplier presentations and business cases. The paper asks what an organisation actually knows about an AI system when the system is liable to change between the day it is tested and the day a decision is taken, and it offers a way to keep deciding without pretending the evidence will hold still.

It opens the Ecaveo strand of these working papers. The strand makes a single argument, that AI adoption is a changing sociotechnical configuration of people, process, data and tools, and that governance has to keep pace with that configuration as it moves. This first paper deals with the most basic consequence of the argument, which is that organisational knowledge about AI carries an expiry date, and that most of the documents boards rely on do not print one.

Editorial balance. Roughly half of the paper reports and interprets published research and public policy documents, about a quarter is my own analysis of what that material implies for organisations, and the remaining quarter sets out a framework of my own (a decay-rate model for claims about AI, with decision rules that follow from it), offered for use and testing and not yet validated.

Legal and regulatory notice. Parts of this paper describe legislation, regulation, regulator activity or court decisions. Nothing in this document constitutes legal advice. The provision of legal advice sits outside the terms of any engagement with the author or with Ecaveo, and the material is presented to support discussion and further review by qualified advisers. Statements about what a source says are attributed to that source. Where I draw a practical implication that a source does not state, the text says so.

HOW TO READ THE EVIDENCE TAGS

Every substantive claim carries a tag stating what kind of evidence supports it. Two tags are combined where a claim rests on both.

PEER-REVIEWED A journal article or peer-reviewed conference paper.

PREPRINT A working paper, preprint or technical report not yet peer reviewed.

STANDARD A published standard or framework, such as ISO or NIST.

POLICY A government, regulator or court document or action.

INDUSTRY A company announcement, press report, analyst or consultancy report.

SCHOLARLY A scholarly book, edited volume or law journal article.

PRACTICE Practitioner observation, including the author's own.

EVIDENCE GAP The available evidence does not answer the question.

ECAVEO The author's own framework or analysis, proposed rather than proven.

CONTENTS

Contents

FRONT MATTER

A note from the author	4
Executive summary	5

PART ONE · THE MOVING OBJECT

1 The pilot that described a system nobody now uses	8
2 Why evidence about AI decays	10

PART TWO · THE ORGANISATIONAL PROBLEM

3 The pacing problem inside the organisation	13
4 A decay-rate model for AI claims	15

PART THREE · KEEPING DECISIONS CURRENT

5 Versioning as a management discipline	19
6 Decision rules for boards under decaying evidence	21

END MATTER

References	24
About the author	25

ILLUSTRATIVE MATERIAL

Every organisation, person and scenario described as illustrative in this paper is invented. Published research and official documents are cited and numbered, and invented material is never presented as evidence.

Research closed 12 December 2022.
Sources and events are as they stood at that date.

CITATION

Forrest, P. (2022) *Knowledge With an Expiry Date: Evidence, Evaluation and Decision-Making When the AI System Keeps Changing*. Ecaveo Working Papers, AI Adoption: The Higher-Order Issues, Paper 1. Ecaveo, 14 December 2022. Available at: ecaveo.com.

Comments and corrections are welcome and will be acknowledged in later versions.

PAPERS IN THIS STRAND

- **Paper one.** *Knowledge With an Expiry Date. Evidence, Evaluation and Decision-Making When the AI System Keeps Changing.* 14 December 2022.
- Paper two. *The Competence Paradox. When AI Makes the Team Better and the Expert Worse.* 21 February 2023.
- Paper three. *The Unit of Adoption.* Why Organisations Adopt Configurations, Not Models. 31 March 2023.
- Paper four. *Everybody's System, Nobody's Risk.* Accountability Gaps in Distributed AI Governance. 28 April 2023.
- Paper five. *Approved Once, Changed Daily.* From Point-in-Time Ethics to Continuous AI Assurance. 31 May 2023.
- Paper six. *The Cost of Waiting.* Precaution, Opportunity and the Ethics of Not Adopting AI. 29 June 2023.

A NOTE FROM THE AUTHOR

Why I have stopped trusting undated conclusions

Much of my advisory work takes place in rooms where a decision is about to be made on the strength of a document someone worked hard to produce. The document is usually careful. It sets out what was tested, what happened and what the organisation should do next, and it rarely says how long any of that will remain true.

That gap did not trouble me much while the systems in question were rules engines or statistical models that changed only when someone in-house changed them. It troubles me more now. Hosted AI services are updated on the supplier's timetable, and the people using them alter how they work as they grow used to the tool, so the thing that was tested and the thing that is deployed drift apart without anyone signing off the difference. In the fortnight since ChatGPT opened to the public I have watched capable people form firm views about what "AI can do" after an afternoon of use, and I have wondered how many of those views will survive the next update.

Coaching has taught me something that transfers. An assessment of a young athlete is a snapshot of a person who is still changing, and the better coaches I know date their judgements and revisit them without embarrassment. I would like boards to treat evidence about AI in the same spirit. It is mostly a habit of mind, although the paper tries to give that habit some structure.

The argument is unfinished, hence a working paper. The decay-rate model in the fourth chapter is mine, tested only against my own practice, and I expect it to change. I would welcome challenge, particularly from anyone who has tried to keep an AI business case current and found a better way to do it.

Paul Forrest
Founder, Ecaveo
14 December 2022

TWO PAGES THAT CIRCULATE ON THEIR OWN

Executive summary

Every claim an organisation makes about an AI system has a half-life, because the models, data, interfaces and user habits underneath the claim keep moving; the answer is to label evidence by how quickly it decays and to build decisions that expect to be revisited. Most AI evaluation treats its subject as a fixed object, such as a model, a version or a benchmark score, yet organisations buy services that change on the supplier's schedule and are used by people who change their own behaviour as they learn. A pilot result or business case may therefore describe a configuration that no longer exists by the time a board reads it.

This paper proposes that organisations sort their claims about AI by how fast each is likely to go out of date, record enough about each observation to know when it has expired, and attach re-test triggers and reversible commitments to decisions resting on short-lived evidence.

1

Evidence about hosted AI services decays on a timetable the buyer does not control.

OpenAI released ChatGPT on 30 November 2022 as a research preview and said it planned regular model updates.¹ **INDUSTRY** Research on human and AI teams shows that even an update which improves a model's accuracy can break the working sense users have built of when to rely on it.² **PEER-REVIEWED**

3

The pacing problem that regulators face also operates inside organisations.

Collingridge's dilemma holds that a technology is easiest to shape when least is known about it.⁵

SCHOLARLY The same gap between change and oversight recurs across emerging technologies.⁶

SCHOLARLY My own view is that procurement, audit and risk committees face a smaller version of the same dilemma on every AI decision. **PRACTICE**

5

Versioning is a management discipline, and the documentation tools for it already exist.

Model cards and datasheets propose structured records of how a model or dataset was made and what it is fit for.^{8,9} **PEER-REVIEWED** Organisations can adapt the same idea to their own observations, recording which system, which configuration and which date each piece of evidence describes. **ECAVEO**

2

Drift and entanglement are well described in machine learning, but rarely carried into organisational decision-making.

The concept drift literature sets out how the relationship between inputs and outcomes shifts over time and how systems can detect and adapt to it.³

PEER-REVIEWED Work on production machine learning shows that changing one part of such a system tends to change the behaviour of the rest.⁴ **PEER-REVIEWED**

Few business cases I see say how long their conclusions will hold. **PRACTICE**

4

Claims about AI decay at different rates, and can be sorted accordingly.

Findings about a specific system expire fastest, findings about a configuration of people, process and tool last longer, and findings about mechanisms, design principles and governance principles last longer again. **ECAVEO** Mechanism findings, such as the long line of research on how people come to over-rely or under-rely on automation, have held up across decades of changing technology.⁷ **PEER-REVIEWED**

6

Public frameworks are moving towards lifecycle oversight, but none yet tells a board when its evidence has expired.

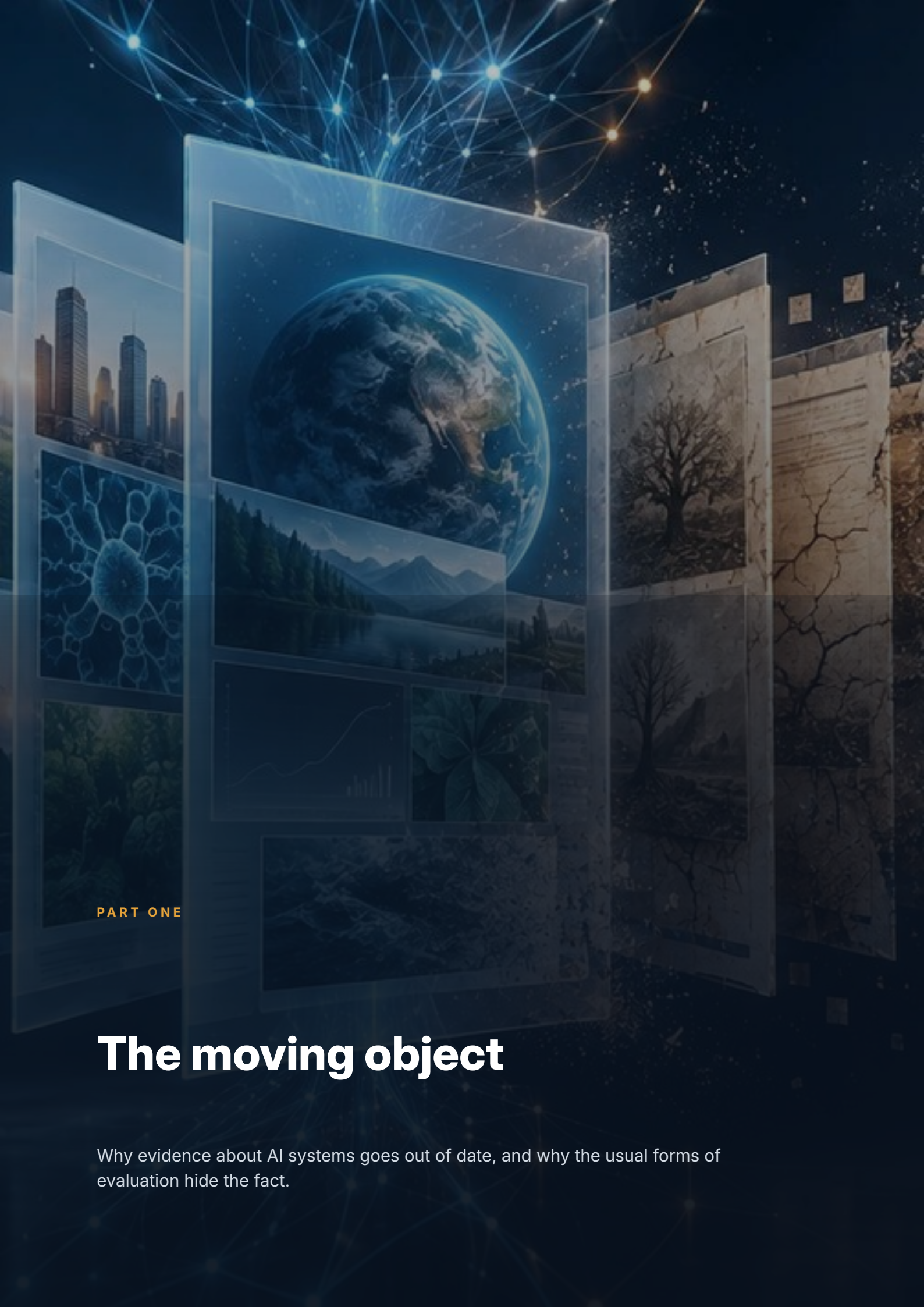
The Council of the EU agreed its general approach on the proposed AI Act on 6 December 2022 and added provisions on general purpose AI.¹⁰ **POLICY** UK, US and NIST documents each stress context, ongoing monitoring or lifecycle risk management.^{11,12,13} **POLICY**

STANDARD How long an evaluation remains valid is left to the organisation. **EVIDENCE GAP**

What follows for a board

A board should expect every AI business case to state what system and configuration its evidence describes, when that evidence was gathered, and what would cause it to be re-tested. It should prefer commitments that can be reversed or scaled down while evidence is young, and it should give a named person the job of deciding when evidence has expired.

None of this requires technical expertise at board level, only that the board asks of AI evidence what it already asks of a financial forecast, which is how old the numbers are and what has changed since.



PART ONE

The moving object

Why evidence about AI systems goes out of date, and why the usual forms of evaluation hide the fact.

CHAPTER ONE

The pilot that described a system nobody now uses

A carefully run pilot can be accurate on the day it ends and misleading by the day it is presented, and nothing in the report will say so.

Organisations have learned to trust pilots. A limited trial, run with real users on real work and measured against a baseline, is the closest most organisations come to an experiment, and the discipline it imposes is valuable. The difficulty with AI is that the pilot tests a moving object and then reports on it as though it had stood still. The report goes to a committee, the committee asks for more detail, the business case goes to the board some weeks later, and at no stage does anyone ask whether the thing that was measured still exists in the form that was measured.

ILLUSTRATIVE SCENARIO**The drafting assistant that changed under the pilot**

A mid-sized regional services organisation runs an eight-week pilot of a hosted text-generation service to draft first replies to customer complaints. The pilot team is careful. It fixes a set of prompts, trains 20 staff, samples drafts for accuracy and tone, and finds that handling time falls while quality holds steady. Three weeks after the pilot closes, the supplier updates the underlying model. Nobody in the organisation is told in terms that mean anything to them, although a line appears in the supplier's release notes. The new model writes longer, warmer replies and handles some complaint types better, but it now occasionally invents a policy reference that the old model never produced. Meanwhile, the pilot staff have learned to skim drafts they have come to trust. The business case, which reaches the board a month later, reports the pilot results accurately, but it describes a configuration of model, prompts and user behaviour that no longer exists.

The scenario is invented, but each of its elements is ordinary. Suppliers of hosted models say openly that they update them. When OpenAI released ChatGPT on 30 November 2022 it called the release a research preview, framed it as part of an iterative approach to deployment, and said it planned to make regular model updates.¹

INDUSTRY Staff who use such a tool daily will, on any reasonable reading of the human factors research, adjust how closely they check its output as their confidence in it grows.⁷ **PEER-REVIEWED** Neither change needs anyone inside the organisation to do anything wrong.

1.1 A gap in the evaluation literature

The research literature on evaluating AI is large, and most of it evaluates a fixed artefact. A model is trained, frozen and scored against a benchmark, and the score is reported as a property of that model. That convention makes sense for science, where results need to be reproducible, but it transfers badly to organisations, which buy ser-

vices that keep changing. The nearest existing work to the problem this paper addresses comes from Bansal and colleagues, who studied what happens when an AI system that people work alongside is updated.² **PEER-REVIEWED** Their central finding is that an update which raises a model's overall accuracy can still lower the performance of the human and AI team, because the new model makes errors in places where the old one was reliable and the user's learned sense of when to trust it no longer fits. They propose measuring the compatibility of an update with the previous version, alongside its accuracy.

What remains thin, as far as I can find, is any account of what happens to organisational knowledge about an AI system when the system underneath it moves.

EVIDENCE GAP By organisational knowledge I mean the business case, the risk assessment, the data protection assessment, the pilot result and the procurement evaluation, all the documents on which a decision rests and to which an auditor or regulator might later turn. These documents are written as statements of fact and filed as though they were permanent.

1.2 Why this is a board matter

It would be convenient to treat all this as a technical concern for the data team. That would be a mistake, for a plain reason. The decision to adopt, extend or withdraw an AI system is taken by executives and boards on the strength of documents, and if those documents describe a system that has since changed, the board is deciding on stale information without knowing it. In financial reporting no board would accept a forecast without knowing its date and assumptions. The same expectation has not yet reached AI evidence, and my own view is that the gap is mostly one of habit. **PRACTICE**

The rest of this part explains why the decay happens. The second part turns to what it does inside organisations and offers a way of sorting evidence by how quickly it expires, and the third part sets out practices and decision rules, closing with questions a board can ask of any AI business case.

COMPATIBILITY

Bansal and colleagues describe an update as compatible when the new model stays correct on the cases where the old model was correct. An update can be more accurate overall and still incompatible, and it is the incompatible errors that catch users out.

CHAPTER TWO

Why evidence about AI decays

Several separate forces wear evidence down, and they compound, so an organisation that watches only one of them will be surprised by the others.

It helps to separate the sources of change, because they call for different responses and are owned by different people. Some sit with the supplier, some with the data, some with the interface and some with the users themselves, and the organisation often has visibility of only one or two.

2.1 Foundation models as moving platforms

The systems that most organisations will meet over the next few years are built on what Bommasani and colleagues termed foundation models, meaning large models trained on broad data that can be adapted to a wide range of downstream tasks.¹⁴

PREPRINT Their report stresses two properties that matter here. The first is homogenisation, meaning that many applications come to rest on the same few underlying models, so that a change or defect in one model passes into everything built on it. The second is emergence, meaning that a model's capabilities are induced implicitly by training and are often discovered only once the model exists. Taken together, these properties imply that the behaviour an organisation observes in its own application depends on decisions made upstream, by people it has never met, on a schedule it does not see.

The range of ways such models can cause harm is also wide. Weidinger and colleagues at DeepMind set out a taxonomy of risks from language models across six areas, running from discrimination and toxic output through information hazards and misinformation to malicious use, harms arising from human and computer interaction, and wider effects on automation, access and the environment.¹⁵ **PREPRINT** For the argument of this paper, what matters is that a change to a model can move its behaviour on several of these dimensions at once, and an evaluation that checked only one of them will not reveal movement on the others.

2.2 Drift in the data and the world

Even a model that never changes will see its performance shift, because the world it is applied to changes. The machine learning literature calls this concept drift, and Lu and colleagues give a thorough review of it.³ **PEER-REVIEWED** They describe drift that arrives suddenly, gradually, incrementally or in recurring patterns, and they organise the response into three activities, which are detecting that drift has occurred, understanding when, how and where it occurred, and adapting the model to it.

For an organisation, drift has a practical meaning. A pilot run on last quarter's customer queries, or on documents written under last year's policy, tells the organisation how the system performed on that material. It does not say how it will perform on the

CONCEPT DRIFT

Concept drift is a change over time in the statistical relationship between a model's inputs

material that arrives after a new product launch, a change in regulation or a shift in the customer base. The drift literature was developed largely for models that organisations train and run themselves, where monitoring is at least possible. With a hosted service the organisation may have no access to the model's internals, and must rely on watching outputs.

and the outcome it predicts. A model trained before the change can go on producing confident outputs that are no longer right.

2.3 Entanglement, interfaces and users

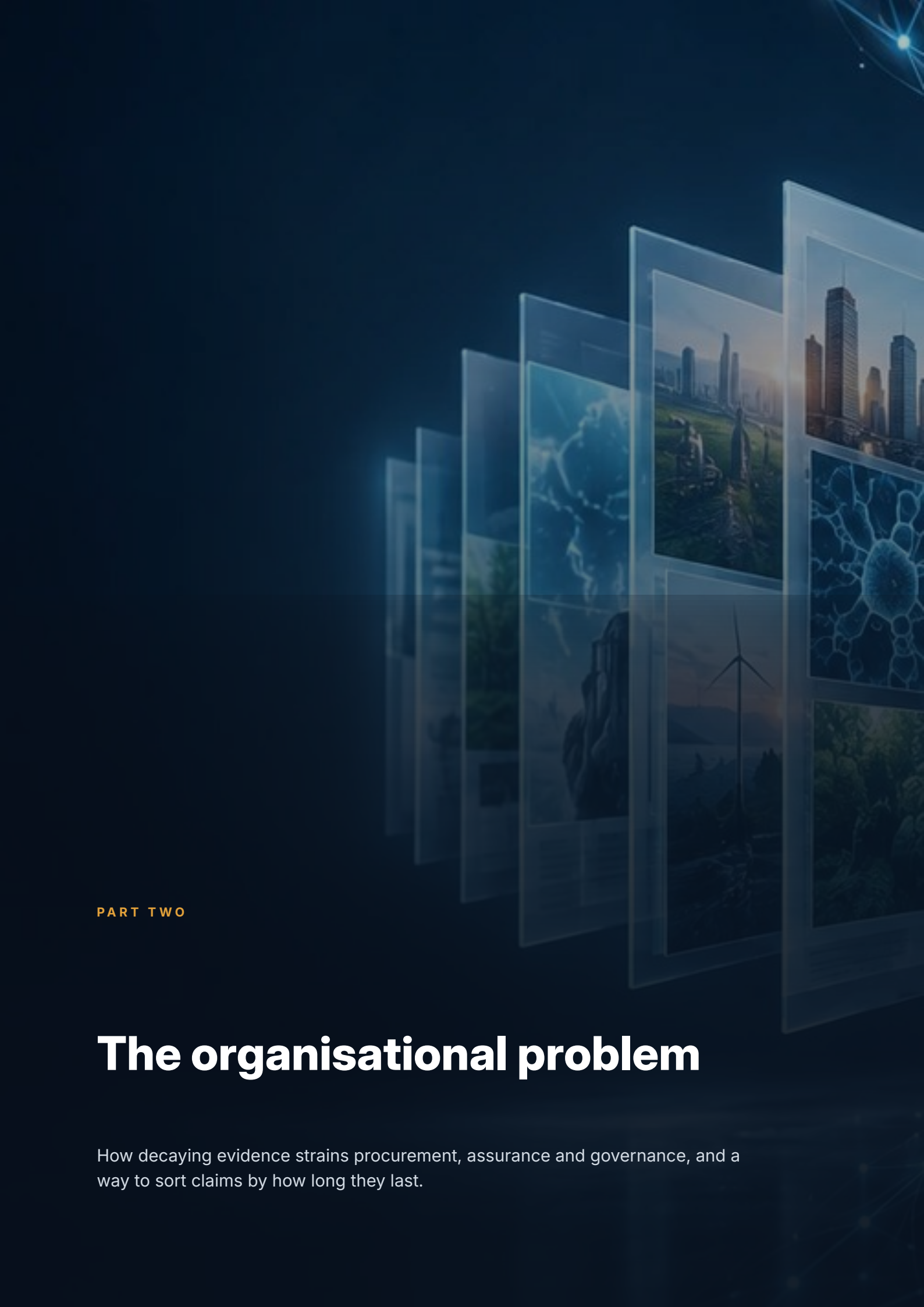
Sculley and colleagues, writing from experience of production systems at Google, argued that machine learning systems have a special capacity for incurring technical debt, the accumulated cost of shortcuts that must be repaid later in maintenance.⁴ **PEER-REVIEWED** Among the problems they identified is entanglement, which they summarised in the principle that changing anything changes everything, along with undeclared consumers of a model's outputs and hidden feedback loops in which a model's outputs shape the data it later learns from. Their paper concerns engineering teams, but the lesson carries over. A change to a prompt template, a retrieval source, a user interface or a downstream workflow can alter what the system does as surely as a change to the model itself.

Users are the last and least documented source of change. The human factors literature has long distinguished appropriate use of automation from misuse, meaning overreliance, and disuse, meaning neglect of automation that would have helped.¹⁶ **PEER-REVIEWED** Lee and See argued that the aim is appropriate reliance, where trust is calibrated to what the automation can actually do, and they described how trust develops through experience and can drift away from the system's true capability.⁷ **PEER-REVIEWED** A pilot population that starts out sceptical and checks every output will produce different results from the same population six months later, when checking has become cursory. Nothing about the model need have changed.

Source of change	Who usually controls it	How visible it is to the buyer
Model update	Supplier	Low, unless contractually notified
Data and world drift	Nobody directly	Moderate, if outputs are monitored
Prompt, retrieval or configuration change	Buyer or integrator	High, if change control applies
Interface or workflow change	Buyer or supplier	Moderate
User adaptation and reliance	Users	Low, unless deliberately studied

TABLE 1 Five sources of change that wear down evidence about an AI system. Author's analysis drawing on Bommasani et al. (2021), Lu et al. (2019), Sculley et al. (2015) and Lee and See (2004).

The table makes one pattern plain. The two sources least visible to the buyer, supplier updates and user adaptation, are also the two that pilots are least designed to capture, since a pilot is usually too short to see either happen.



PART TWO

The organisational problem

How decaying evidence strains procurement, assurance and governance, and a way to sort claims by how long they last.

CHAPTER THREE

The pacing problem inside the organisation

The dilemma that governments face in regulating new technology repeats, in miniature, every time an organisation approves an AI system.

Discussion of the pacing problem has mostly concerned public policy. Organisations have their own version of it, and they have given it much less thought.

3.1 Collingridge's dilemma, brought indoors

David Collingridge set out the dilemma that bears his name in 1980.⁵ **SCHOLARLY** Early in a technology's life, change is easy but too little is known about its consequences to know what to change. Later, once the consequences are visible, the technology is embedded in practice, investment and expectation, and change has become expensive and slow. It is a problem of information first and of power afterwards. Marchant, Allenby and Herkert gathered work showing the same structure across a range of emerging technologies, in which legal and ethical oversight struggles to keep up with the speed of technical change.⁶ **SCHOLARLY**

The dilemma translates directly into organisational life, and the translation is my own inference beyond what either source claims. **PRACTICE** A procurement team evaluating an AI service early has little evidence about how it will behave in the organisation's own context. By the time that evidence exists, the service is integrated into workflows, staff have been trained, a contract has been signed and savings have been counted in a budget. Withdrawing it has become costly. The information problem and the power problem arrive in the same order they do for governments, only faster and with fewer people watching.

3.2 Where assurance processes assume stability

Most organisational assurance was designed around systems that change through controlled releases. A risk assessment is completed at approval and reviewed once a year, and a data protection impact assessment is written before processing begins. An internal audit tests controls at a point in time and reports. Each of these practices assumes, reasonably enough for conventional software, that the system under review will stay recognisably the same until someone in the organisation decides to change it.

Hosted AI breaks that assumption in two places. The supplier can change the model without any internal change request, and users can change their reliance on it without any change to the system at all. An annual review may therefore certify a configuration that has changed several times since the last one, and the certification will read as current. I have not found published evidence on how often this happens in practice, and it would be useful to have some. **EVIDENCE GAP**

3.3 The regulators are moving too

The pacing problem also runs in the other direction, since the rules organisations must work under are themselves in motion. The European Commission proposed its AI Act in April 2021.¹⁷ **POLICY** The Council of the EU agreed its general approach on 6 December 2022, some 20 months later.¹⁰ **POLICY** According to the Council's announcement, the text narrows the definition of an AI system and adds provisions for general purpose AI systems, with the requirements for those systems to be specified through an implementing act, and negotiations with the European Parliament can begin once the Parliament has adopted its own position. Systems of the kind that ChatGPT represents were not the focus of the original proposal, which gives some sense of how far the technology has moved while the legislation was being drafted.

The UK government set out in July 2022 its intention to take a context-specific, pro-innovation approach, in which existing regulators apply a set of cross-sectoral principles within their own domains, and it said a white paper would follow.¹¹ **POLICY** In the US, the White House Office of Science and Technology Policy published its *Blueprint for an AI Bill of Rights* in October 2022, a non-binding document whose principles include safe and effective systems supported by pre-deployment testing and ongoing monitoring.¹² **POLICY** NIST released a second draft of its AI Risk Management Framework in August 2022. The draft is organised around governing, mapping, measuring and managing risk across the lifecycle of a system.¹³ **STANDARD**

My reading, as a business adviser and not a lawyer, is that these documents share a direction. Each treats oversight as something that continues after deployment. None, as far as I can see, tells an organisation how long a given evaluation may be relied upon, and that judgement appears to be left with the organisation. **EVIDENCE GAP** Boards should expect the requirements to change again before most current AI projects reach maturity, and should take advice on the detail from qualified advisers.

A decision taken on young evidence should be built to be revisited. Collingridge's own prescription, broadly, was to keep technologies correctable while the facts are still coming in. For an organisation that means favouring commitments that can be reversed, scaled back or re-tendered, and naming the evidence that would trigger a second look before the first decision is signed.

CHAPTER FOUR

A decay-rate model for AI claims

Claims about AI do not all expire at the same speed, so sorting them by how quickly they decay lets an organisation keep its durable knowledge and refresh the rest.

If evidence about AI decays, the obvious responses are to stop relying on it or to re-test everything constantly. Neither is practical. The first leaves the organisation deciding on instinct, and the second consumes more effort than most adoption projects can bear. A more useful response starts from the observation that different kinds of claims decay at different rates, and that an organisation can say in advance roughly which kind of claim it is making. **ECAVEO**

4.1 Five tiers of durability

The model I propose sorts claims into five tiers, from the fastest-decaying to the slowest. **ECAVEO** The tiers are defined by what a claim is about.

1	System-specific findings	Claims about one product or model version, such as its error rate on our data.
2	Configuration findings	Claims about a mix of tool, prompts, process and people in one setting.
3	Mechanism findings	Claims about how people and systems interact, such as how reliance forms.
4	Design principles	Claims about what good design of an AI-supported process requires.
5	Governance principles	Claims about how decisions, accountability and oversight should work.

FIGURE 1 Five tiers of claims about AI, from fastest to slowest decay. Source: author's framework.

System-specific findings are the most concrete and the most perishable. "The service misclassified four per cent of our test invoices" is a claim of this kind, and it describes a particular model version on a particular sample at a particular time. It may be invalid after the next supplier update, and it says little about any other product.

Configuration findings describe a combination of a tool, the prompts or settings around it, the process it sits in and the people who use it. "Complaint handlers using the assistant with a mandatory review step cut handling time without a rise in upheld complaints" is such a claim. It can survive a minor model update. It will not survive a change to the review step, and it may not survive the handlers becoming more trusting.

Mechanism findings describe how things work across many systems. The finding that people tend to over-rely on automation that has performed well, and to under-rely on automation after seeing it fail, is a mechanism claim, and it has held across decades of changing technology.^{16,7} **PEER-REVIEWED** Bansal and colleagues' finding that accuracy gains can come with compatibility losses is another.² **PEER-REVIEWED** Such claims tell an organisation what to watch for in any system it adopts.

Design principles and governance principles are the most durable. "Place human review where the model's errors are costliest, and check that reviewers are actually reviewing" is a design principle. "Every AI-supported decision should have a named owner who can explain and reverse it" is a governance principle. These outlast many generations of technology because they concern how the organisation itself works, and organisations change more slowly than their tools.

Tier	Illustrative claim	Plausible shelf life	Typical re-test trigger
System-specific	Error rate of model version X on our invoices	Weeks to months	Any supplier model update
Configuration	Handling time falls with assistant plus review step	Months	Change to prompts, process, staff or model
Mechanism	Users check less as confidence grows	Years	New evidence from research
Design principle	Put review where errors cost most	Years	Change in the organisation's risk appetite
Governance principle	Named owner for every AI-supported decision	Many years	Change in law or board structure

TABLE 2 The decay-rate model, with illustrative claims and triggers. Source: author's framework; shelf lives are judgements for discussion, not measured values.

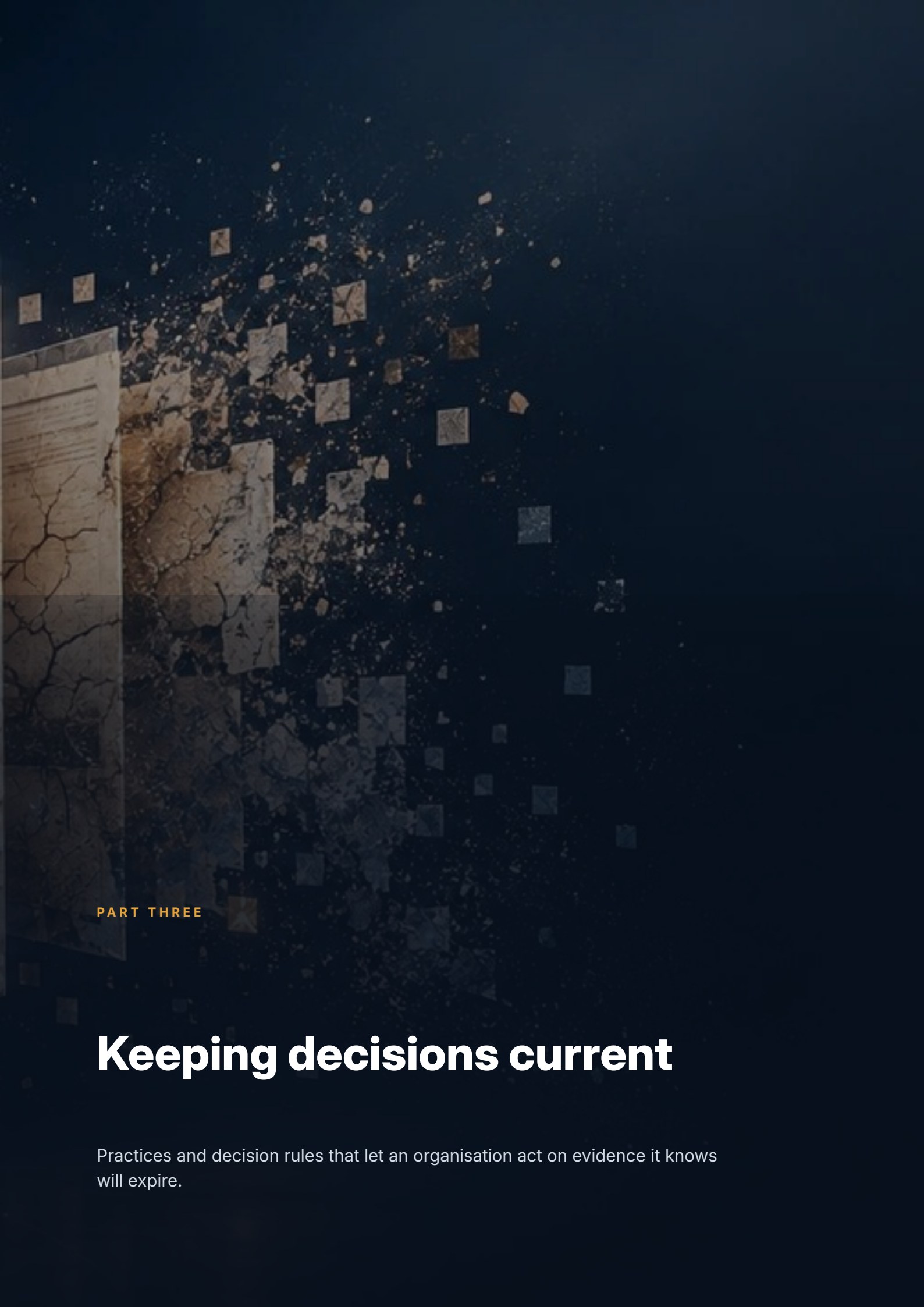
4.2 Using the tiers

The tiers do two jobs. The first is labelling. When a business case states that "the tool is accurate", a reader should be able to ask which tier the statement belongs to, and the answer usually reveals that a system-specific finding is being presented with the confidence of a principle. The second job is allocation of effort. An organisation can afford to re-test system-specific and configuration findings frequently only if it is not also re-litigating its principles every quarter, so the model lets it hold the slow tiers steady while refreshing the fast ones. **ECAVEO**

The tiers also clarify what a pilot can and cannot establish. A well-run pilot produces mainly configuration findings, some system-specific findings and, if it is designed for it, a little evidence on mechanisms. It does not validate a design principle, though it can illustrate one, and it says almost nothing about a different supplier's product. My own impression from reading business cases is that the boundaries between these tiers are routinely blurred in the step from pilot report to board paper, often with no intent to mislead. **PRACTICE**

4.3 Limits of the model

The model is a heuristic and should be read as one. The shelf lives in the table are judgements, and I know of no study that measures how long a pilot finding about a hosted AI service remains decision-useful. **EVIDENCE GAP** The boundaries between tiers are also porous (a mechanism finding can turn out to be an artefact of one generation of technology, and a governance principle can be overturned by a single piece of legislation). There is a deeper question I have not resolved, which is whether the tiers decay at independent rates or whether a large enough change at the system tier, of the kind foundation models may bring, could reach up and unsettle the tiers above it. I suspect it could, and that this is where the model will need most work.



PART THREE

Keeping decisions current

Practices and decision rules that let an organisation act on evidence it knows will expire.

CHAPTER FIVE

Versioning as a management discipline

An observation that does not record what it observed, and when, cannot be known to have expired.

Software engineers learned long ago to version their code. Organisations using AI need an equivalent habit for the observations they make about systems whose internals they often cannot see.

5.1 What to record

The documentation research offers a starting point. Mitchell and colleagues proposed model cards, short documents that accompany a trained model and report its intended use, the conditions under which it was evaluated, and its performance across relevant groups.⁸ **PEER-REVIEWED** Gebru and colleagues proposed datasheets for datasets, which record why a dataset was created, what it contains, how it was collected and what uses it suits.⁹ **PEER-REVIEWED** Both proposals rest on the same idea, that a technical artefact should travel with a structured account of where it came from and what it is fit for.

Organisations can adapt that idea to their own evidence without waiting for suppliers to adopt it. **ECAVEO** Every observation that feeds a decision should carry a short record of what was observed.

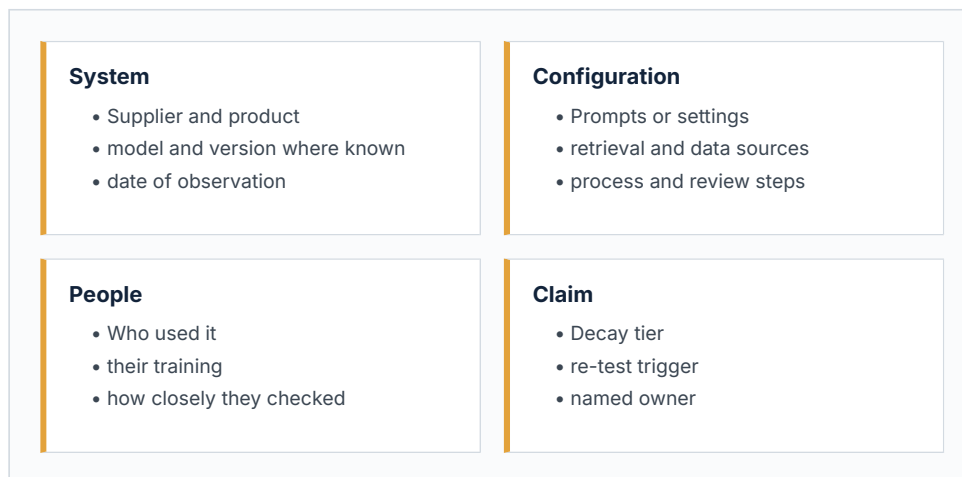


FIGURE 2 What an evidence record should capture for each observation of an AI system. Source: author's framework, adapted from Mitchell et al. (2019) and Gebru et al. (2018).

Where the supplier does not disclose a model version, the record should say so, since that is itself a finding about how quickly the evidence may decay. The date of observation matters more than it seems. With a hosted service, the date may be the only reliable proxy for the version.

5.2 Asking suppliers for change information

A record of the organisation's own observations is only half the picture if the supplier can change the system without notice. Contracts are one lever. Procurement contracts can, in principle, ask suppliers to notify material changes to the models underlying a service, to maintain release notes that a non-specialist can read, or to allow a period in which a previous version remains available. Whether suppliers of general-purpose models will agree to such terms, and whether they can define "material" in a way that means something to a buyer, is an open question. **EVIDENCE GAP** My view is that buyers should ask anyway, because the asking reveals how the supplier thinks about change, and because the same request made by many buyers tends to shape what suppliers offer. **PRACTICE**

5.3 Watching the users as well as the system

Versioning the system is not enough when the users are a source of change. The human factors evidence suggests that reliance shifts with experience, so an organisation that wants to know whether its review step still works needs to look at reviewing behaviour over time, for instance by sampling reviewed outputs for errors that the reviewer passed.⁷ **PEER-REVIEWED** This is cheap to do and seldom done. It also gives the organisation some protection against the incompatible update Bansal and colleagues describe, since a rise in errors that slip past review is often the first sign that the model's failure pattern has moved.² **PEER-REVIEWED**

CHAPTER SIX

Decision rules for boards under decaying evidence

A board can decide well on imperfect evidence, provided it knows how old that evidence is and what would make it look again.

This final chapter sets out my recommendations and then closes with the questions this paper leaves open. The recommendations are my own, and I offer them for use and testing. **ECAVEO**

6.1 Three decision rules

The first rule is that every piece of evidence in an AI business case should carry an expiry condition. That condition may be a date, an event such as a supplier model update, or a threshold such as a rise in reviewed errors, and it should be written into the paper the board approves. When the condition is met, the evidence is treated as expired until re-tested. Any decision resting on it is then flagged for review. **ECAVEO**

The second rule is that re-test triggers should be defined before approval. A trigger set afterwards tends to be set in the light of results the organisation hopes to keep. Useful triggers include a notified or detected change of model, a change to prompts, retrieval sources or review steps, a material change in the population of cases the system handles, a change in the relevant regulatory requirements, and signs that reliance has drifted, such as falling review times without a matching fall in error rates. **ECAVEO**

The third rule is that, while evidence is young, commitments should be reversible. In practice that means shorter initial contract terms or break clauses, retaining the ability to run the previous process for a period, staged roll-outs that can be paused, and budgets that do not bank savings before configuration findings have survived at least one re-test. This is Collingridge's prescription in a corporate setting.⁵ **ECAVEO**

1	Record	Log each observation with system, configuration, people, date and tier.
2	Decide	Approve with expiry conditions and re-test triggers written in.
3	Commit reversibly	Prefer break clauses, staged roll-out and a retained fallback.
4	Monitor	Watch supplier changes, data drift and reviewer behaviour.
5	Re-test	When a trigger fires, treat evidence as expired and test again.

FIGURE 3 A decision loop for AI adoption under decaying evidence. Source: author's framework.

6.2 Questions a board can ask

A board does not need to master the technology to apply these rules. It needs to ask a small set of questions of every AI business case and to expect specific answers.

ECAVEO

1. Which system, version and configuration does this evidence describe, and when was it gathered?
2. For each main claim, is it about this system, this configuration, a general mechanism or a principle?
3. What would cause this evidence to expire, and who will notice when that happens?
4. What has the supplier committed to tell us about changes to the underlying model?
5. How are we checking that the people reviewing outputs are still reviewing them?
6. If the evidence turns out to be wrong, how quickly and at what cost can we reverse this decision?

The last question matters most while evidence is young. A board that receives a confident answer to the first five and a vague answer to the sixth should, in my view, approve a smaller or more reversible version of the proposal. **PRACTICE**

It also helps to name an owner. Someone inside the organisation has to decide that evidence has expired, and the role sits awkwardly between the business sponsor, who wants the project to continue, and the risk function, which may lack the technical access to judge. I would give the role to the business sponsor, with a duty to report expiry events to the risk committee, but I hold that view loosely and can see arguments for placing it elsewhere. **ECAVEO**

6.3 Conclusion and open questions

The argument of this paper is simple to state. Evidence about AI systems has a half-life, because models, data, configurations and users all change, and much of that change happens outside the buyer's control and out of its sight. The appropriate response is to label evidence by how quickly it decays, to record what each observation describes, and to build decisions that expect to be revisited. None of this makes the evidence permanent, and it is not meant to. It makes the impermanence visible, which is the precondition for governing it.

Several questions remain open. The working papers that follow in this strand will return to some of them. How long, in practice, does a pilot finding remain decision-useful for a hosted AI service, and does the answer differ by use case? Who should own the judgement that evidence has expired, and how is that owner held to account? Can procurement contracts oblige suppliers of general-purpose models to notify material changes, and would suppliers be able to say what "material" means? And, less tract-

ably, does a large enough change in the underlying technology unsettle the durable tiers as well as the fragile ones, so that even principles need an expiry condition of their own?

END MATTER

References

1. OpenAI (2022) ChatGPT: optimizing language models for dialogue. Blog post, 30 November 2022. Available at: openai.com/blog/chatgpt.
2. Bansal, G., Nushi, B., Kamar, E., Weld, D.S., Lasecki, W.S. and Horvitz, E. (2019) Updates in human-AI teams: understanding and addressing the performance/compatibility tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(01), 2429-2437.
3. Lu, J., Liu, A., Dong, F., Gu, F., Gama, J. and Zhang, G. (2019) Learning under concept drift: a review. *IEEE Transactions on Knowledge and Data Engineering* 31(12), 2346-2363.
4. Sculley, D. et al. (2015) Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems* 28.
5. Collingridge, D. (1980) *The Social Control of Technology*. London: Frances Pinter.
6. Marchant, G.E., Allenby, B.R. and Herkert, J.R. (eds) (2011) *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem*. Dordrecht: Springer.
7. Lee, J.D. and See, K.A. (2004) Trust in automation: designing for appropriate reliance. *Human Factors* 46(1), 50-80.
8. Mitchell, M. et al. (2019) Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT 2019)**, 220-229.
9. Gebru, T. et al. (2018) Datasheets for datasets. arXiv:1803.09010. Also published in *Communications of the ACM* 64(12), 86-92, 2021.
10. Council of the EU (2022) Artificial Intelligence Act: Council calls for promoting safe AI that respects fundamental rights. Press release, 6 December 2022.
11. DCMS, BEIS and Office for AI (2022) *Establishing a pro-innovation approach to regulating AI*. Policy paper, 18 July 2022.
12. White House OSTP (2022) *Blueprint for an AI Bill of Rights*. 4 October 2022.
13. NIST (2022) *AI Risk Management Framework: Second Draft*. National Institute of Standards and Technology, 18 August 2022.
14. Bommasani, R. et al. (2021) On the opportunities and risks of foundation models. arXiv:2108.07258.
15. Weidinger, L. et al. (2021) Ethical and social risks of harm from language models. arXiv:2112.04359.
16. Parasuraman, R. and Riley, V. (1997) Humans and automation: use, misuse, disuse, abuse. *Human Factors* 39(2), 230-253.
17. European Commission (2021) *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. COM(2021) 206 final, 21 April 2021.

END MATTER

About the author

Paul Forrest is the founder of Ecaveo. He advises boards and leadership teams on the adoption and governance of artificial intelligence, drawing on a background in business process reengineering and on hands-on work building machine learning systems, including small, domain-specific language models. He has served as a special adviser to UK government on the Global Entrepreneur Programme and works as a board member and business adviser.

He is also a qualified athletics coach who coaches sprint, hurdle and endurance athletes as a volunteer, and much of his thinking about how people learn, decide and improve comes from the track.

This paper is part of the Ecaveo Working Papers, AI Adoption: The Higher-Order Issues. The views expressed are the author's own.

ecaveo.com

Knowledge With an Expiry Date

Ecaveo Working Papers, AI Adoption: The Higher-Order Issues
Paper one of six · 14 December 2022