



Approved Once, Changed Daily

From Point-in-Time Ethics to Continuous AI Assurance

About this paper

This paper is written for board members, risk and compliance leads, and the chairs of approval committees and research ethics panels who are asked to sign off systems that use machine learning. It addresses one practical problem. Most of the approval instruments these groups rely on were designed for a product, process or study that could be described at the outset and would stay largely as described, and adaptive AI systems do not behave that way once they are in use.

The paper is the fifth of six in the Ecaveo working papers on the higher-order issues in AI adoption. That strand argues that AI adoption is a changing sociotechnical configuration whose governance has to keep pace with it as it changes. This paper applies the argument to approval.

Editorial balance. Roughly two fifths of the paper reports published research, standards and regulatory documents, about a third is my analysis of what that material means for approval bodies, and the remainder, concentrated in the last three chapters, sets out a continuous assurance model of my own that is proposed and has not been tested.

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PREPRINT A working paper, preprint or technical report not yet peer reviewed.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

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.

CONTENTS

Contents

FRONT MATTER

A note from the author	4
Executive summary	5

PART ONE · THE APPROVAL PROBLEM

1	The system that was approved	9
2	Why point-in-time approval fails	11

PART TWO · WHAT THE FIELD ALREADY KNOWS

3	From principles to continuing assessment	15
4	What the frameworks and regulators expect	17

PART THREE · A WORKING MODEL

5	A continuous assurance model	21
6	Proportionality and research ethics	24
7	Recommendations and open questions	26

END MATTER

References	28
About the author	29

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 30 May 2023. Sources and events are as they stood at that date.

CITATION

Forrest, P. (2023) *Approved Once, Changed Daily: From Point-in-Time Ethics to Continuous AI Assurance*. Ecaveo Working Papers, AI Adoption: The Higher-Order Issues, Paper 5. Ecaveo, 31 May 2023. 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

The folder nobody reopens

Over years of advisory work I have advised, reported to and occasionally sat on approval committees, and one pattern recurs often enough that I no longer put it down to bad luck. A proposal arrives with a careful pack, the committee asks good questions, conditions are set, and the minutes record approval. The pack then goes into a folder. When somebody next opens that folder it is usually because something has gone wrong, and the first discovery is that the system described in the pack differs from the system that went wrong.

For most of the technology I have seen approved, that gap was tolerable. A payroll platform changes, but it changes through releases that somebody plans and somebody else tests, and the approval can be refreshed at those moments if anyone cares to. Systems built on machine learning have loosened that link. Their behaviour can shift because the data feeding them shifts, because a supplier updates a hosted model, or because a team edits a prompt on a Tuesday afternoon without regarding the edit as a change at all.

I wrote this paper because I kept giving the same advice in slightly different words, and because the advice was harder to follow than it sounded. "Keep the assessment under review" is easy to minute and hard to do. What committees lack, in my experience, is a small set of concrete commitments that make review happen without turning every model update into a fresh submission. (The committees I have watched try the fresh-submission route tired of it quickly, which says something about where the burden belongs.)

The framework in the later chapters is mine and untested. I offer it as a starting point for committees willing to try it and to report back on where it fails, since failure reports are the evidence this field most lacks.

Paul Forrest
Founder, Ecaveo
31 May 2023

TWO PAGES THAT CIRCULATE ON THEIR OWN

Executive summary

An approval given once to an adaptive AI system certifies a system that will soon cease to exist, and the answer is a small, standing set of commitments that tell an organisation when to look again, not a heavier approval at the start.

Ethics review, data protection impact assessment and algorithmic impact assessment were all built around an object that can be described at the outset, whereas machine learning systems change after sign-off through supplier updates, edited prompts, new data and new uses. Each of those changes can alter who is affected and how. Research has moved from principles to tools and on to auditing, and the main risk frameworks now treat risk management as continuing across the lifecycle.¹ **STANDARD** What remains thin, outside regulated medical devices, is the operational bridge that would let an approval committee, a research ethics panel or a risk function govern a system that keeps changing. This paper proposes one.

1

Machine learning systems change in ways that ordinary change control does not see.

Sculley and colleagues argued that production machine learning accumulates hidden technical debt, because changing any one input can change the behaviour of the whole system.² **PEER-REVIEWED** The concept drift literature shows that learned relationships can move over time.³ **PEER-REVIEWED** A May 2023 preprint reports that training generative models on generated content erases the tails of the original distribution, an effect it calls model dementia.⁴

PREPRINT

3

Audit and impact assessment are increasingly framed as continuing work, but the methods are young.

Raji and colleagues set out an end-to-end framework for internal algorithmic audit,⁷ **PEER-REVIEWED** and Minkinen, Laine and Mäntymäki have conceptualised continuous auditing of AI.⁸ **PEER-REVIEWED** Metcalf and colleagues argue that impacts are co-constructed during assessment, which suggests that an assessment fixed at one date fixes one construction of harm.⁹ **PEER-REVIEWED** How continuous methods perform in ordinary organisations is undocumented.

EVIDENCE GAP

5

Medical device regulation offers the most developed model for governing change after approval.

The US Food and Drug Administration (FDA) proposed in 2019 that manufacturers specify anticipated modifications in advance.¹² **POLICY** Its draft guidance of April 2023 develops that idea into a predetermined change control plan, reviewed at submission, so that each described modification does not need a submission of its own.¹³ **POLICY**

2

Principles and one-off tools have not closed the distance between approval and practice.

Mittelstadt argued that principles alone cannot guarantee ethical AI, partly because AI development lacks the professional structures that give medical ethics its force.⁵ **PEER-REVIEWED** Morley and colleagues found the available tools for putting principles into practice spread unevenly across principles and across stages of development.⁶ **PEER-REVIEWED**

4

The main frameworks and the UK white paper already expect risk to be managed across the lifecycle.

The NIST framework says risk management should be continuous and calls for post-deployment monitoring plans that cover incidents, decommissioning and change management.¹ **STANDARD** ISO/IEC 23894 gives guidance on managing AI risk that applies across the life of a system,¹⁰ **STANDARD** and the government's white paper of 29 March 2023 expects AI risks to be identified, assessed and managed continually throughout the AI lifecycle.¹¹ **POLICY**

6

A proportionate continuous assurance model can be assembled from nine plain elements.

The model proposed here combines a versioned system description, a change log, agreed reassessment triggers, performance monitoring, bias monitoring, an incident and near-miss route, consent renewal, supplier change ownership, and retirement and rollback criteria. **ECAVEO** Its demands scale with the system's effect on people, so a small organisation can apply it.

ECAVEO**What follows for a board**

A board should stop treating AI approval as a single event. For each system above a modest impact threshold it should be able to say who owns the current description of the system, which changes would send it back for review, what evidence of its behaviour in use arrives and how often, and who hears about it when a supplier changes the model underneath it.

None of this requires a large function, only conditions of approval written as checkable commitments with a named person against each. My own view is that the thorough, once-only approval has had its day for these systems, and that proportionate, light and ongoing assurance will serve the people they affect far better.



PART ONE

The approval problem

Why a one-off sign-off cannot hold for systems that keep changing after it is given.

CHAPTER ONE

The system that was approved

An approval certifies a description of a system. With adaptive AI the description and the running system drift apart, and most approval processes have no means of noticing when they do.

Approval committees, research ethics panels and data protection functions work from an assumption they seldom state. They take it that the thing they are asked to assess can be described at the start, and that it will stay close enough to that description for the assessment to hold until somebody brings it back. For most of the history of these bodies the assumption has served well, because a study protocol or a conventional software system changes through amendments and releases that a named person initiates and that can be routed back to whoever gave the approval.

Machine learning systems weaken that link. My argument in this paper is that a one-off approval of an adaptive AI system certifies a system that will soon not exist, and that without defined triggers for re-review the approval turns into a record whose main beneficiary is the organisation's own file. Continuous assurance is the obvious remedy, although the phrase alarms committees who picture a permanent stream of resubmissions, and much of what follows is an attempt to show that it can be light.

1.1 What the committee signed

ILLUSTRATIVE SCENARIO

A repairs triage tool, approved in the spring

An invented housing association approves an AI tool that sorts incoming repair requests by urgency. The approval pack is careful. It names the supplier's model, lists the fields the tool reads and the categories it assigns, reports a pilot on six months of historical requests, and includes a bias check across the association's main tenant groups. The committee approves the tool with conditions, and the minutes record a review in 12 months.

Within the year, and without anyone treating it as a change, four things happen. The supplier moves the tool onto a newer model. A service manager, finding that it seemed to help, adds the call handlers' free-text notes to the fields the tool reads for each request. A neighbouring team starts using the same tool to prioritise complaints, which were never in scope. A cold winter changes the mix of requests, with far more heating failures than the pilot data held. When the review date arrives, the committee receives the original pack with a covering note saying the tool continues to perform well, and approves its continued use. Nobody has been dishonest, and the committee has reviewed a system that no longer exists.

The case is invented, yet each of its moves is ordinary, and I have met versions of most of them in advisory work. **PRACTICE** No single change looked significant to the person who made it. The supplier's upgrade arrived as a release note, the extra field was an

operational improvement, the new use was a colleague being helpful, and the winter was weather. Taken together, they altered the tool's underlying model, its inputs, the population it served and its purpose, which between them account for most of what the committee had assessed.

1.2 Where the failure sits

The committee in the scenario did what it was designed to do, and that is the difficulty. Its instrument was a single assessment with a dated review, which presumes that change arrives through channels the committee controls and at a pace the review date can absorb. Neither presumption held. The review date was a guess, made before anyone knew how the tool would be used, about when it might have changed enough to matter; and the covering note answered a question about performance that nobody had defined in measurable terms at the outset.

This paper belongs to a strand whose single argument is that AI adoption is a changing sociotechnical configuration, meaning an arrangement of models, data, people, suppliers and purposes that shifts over time, and that governance has to keep pace with the arrangement as it shifts. Approval is where that argument meets the committee room most directly. Part one sets out why point-in-time approval fails. Part two reviews what researchers, standards bodies and regulators had said about the problem by the end of May 2023, and Part three proposes a continuous assurance model that an ordinary organisation could adopt, with a section on research ethics panels, whose instruments are tied more tightly than most to a single moment of review.

CHAPTER TWO

Why point-in-time approval fails

Machine learning systems change through more routes than conventional software, and several of those routes never pass through a release process that anyone would recognise.

2.1 Entanglement and hidden debt

Writing from Google in 2015, Sculley and colleagues argued that machine learning brings its own kind of technical debt, one that builds up at the level of the whole system and is hard to see from inside any single component.² PEER-REVIEWED They observed that the learning code is usually a small fraction of a deployed system. Around it sits far larger machinery for collecting and checking data, extracting features, managing configuration, serving predictions and monitoring. They also named a principle they called CACE, short for "Changing Anything Changes Everything", to describe how altering one input, one threshold or one data source can shift the behaviour of the model as a whole.² PEER-REVIEWED

For an approval body the implication is uncomfortable. Conventional change control asks whether the code has changed, whereas Sculley and colleagues point to changes that leave the code untouched, including shifts in upstream data, consumers of a model's outputs who were never declared, and feedback loops in which a model's decisions shape the data it will later learn from.² PEER-REVIEWED In 2017 Breck and colleagues turned part of this analysis into a production readiness rubric, and they gave monitoring of the live model a group of tests of its own alongside tests of data, model and infrastructure.¹⁴ PEER-REVIEWED The reasoning is simple enough, since a model that passed its checks at launch can still fail quietly afterwards.

2.2 Four routes to change

In advisory work I find it useful to separate four routes by which a deployed AI system changes after approval, because each has a different owner and a different chance of being noticed. PRACTICE

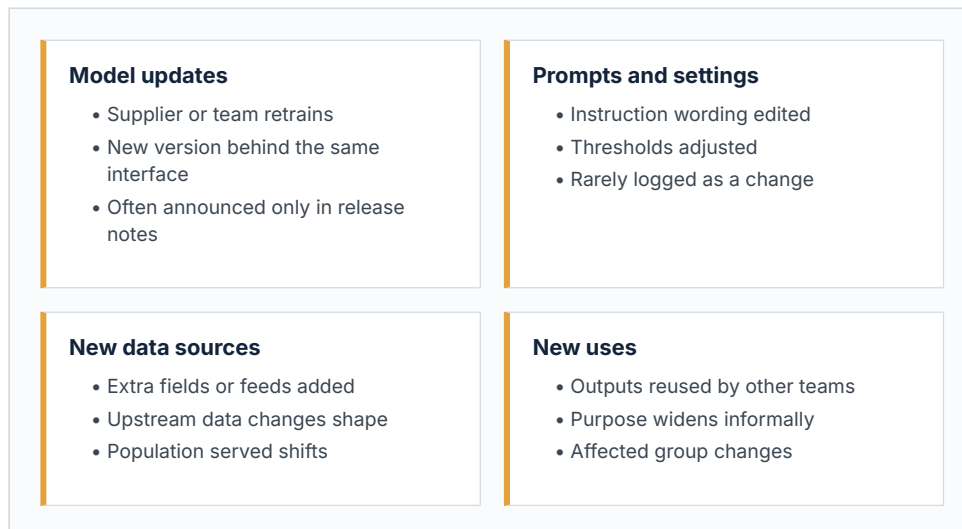


FIGURE 1 Four routes by which an approved AI system changes after sign-off. Source: author's framework, drawing on Sculley et al. (2015).

Model updates are the most visible of the four, although visibility depends on who hosts the model. An organisation that trains its own model at least knows when it re-trains. An organisation that reaches a supplier's model through an interface may learn of a new version from a developer notice written for engineers, if it learns at all, and in my experience those notices rarely reach the committee that approved the use.

PRACTICE

Changes to prompts and settings are the least visible. With tools built on large language models, the wording of the standing instructions given to the model is part of the system, and a small edit can change what the tool produces for every user; yet the edit is typically made by an operational team in the course of ordinary work and is seldom logged as a change. **PRACTICE** New data sources change what the model sees, and often who it sees, which bears directly on any bias check done at approval. New uses are the route most likely to change the ethical question itself, since a tool approved for one purpose and one group of people acquires a different set of affected parties the moment a second team starts relying on it.

Only the first route is routinely recorded anywhere, and even that record usually sits with a supplier. **PRACTICE**

2.3 Drift, degradation and model-generated data

Even when nobody changes anything, the world the model learned from moves on. Gama and colleagues, in their survey of the field, define concept drift as a change over time in the relationship between input data and the target a model is trying to predict, and they distinguish this real drift from changes in the distribution of the inputs alone.³ **PEER-REVIEWED** Their survey treats adaptation as a standing need for any learning system that works on data arriving over time, which describes most systems an approval committee sees. The winter in the scenario is a homely case of the second kind, a change in the mix of requests that may or may not alter what urgency means.

A newer concern arrived just before this paper was completed. On 27 May 2023, Shumailov and colleagues released a preprint reporting that when generative models are trained on content produced by earlier models, the tails of the original content distribution disappear, and they describe the resulting defects as irreversible.⁴

PREPRINT They call the effect model dementia and report it in variational autoencoders, Gaussian mixture models and large language models, arguing that data from genuine human interaction will become more valuable as generated content spreads across the web.⁴ **PREPRINT** The work had not been peer reviewed when this paper closed, so I treat it as a signal worth watching until it has been.

If the finding holds, it gives one more reason to doubt that a model's behaviour will stay stable across successive versions, and it points at the tails of a distribution, which is where rare cases and small groups of people sit. That last connection is my inference and goes beyond what the preprint claims. It matters for approval because a bias check at sign-off is, in effect, a check on how a model treats the tails.

READING PREPRINTS

A preprint is public but has not yet passed peer review. I cite one here because the field moves faster than journals, and I tag it so that readers can weigh it accordingly.



PART TWO

What the field already knows

Research, standards and regulators have each moved towards lifecycle thinking, by different routes and at different speeds.

CHAPTER THREE

From principles to continuing assessment

The ethics literature has moved from stating principles, to building tools, to auditing. Each move has brought assessment closer to the running system without yet reaching it.

3.1 Principles without a method

Mittelstadt observed in 2019 that the many published statements of AI ethics principles had converged on something resembling medical ethics, and then argued that the resemblance was misleading.⁵ **PEER-REVIEWED** Medicine, on his account, has four things AI development lacks, namely common aims and fiduciary duties, professional history and norms, proven methods for translating principles into practice, and robust legal and professional accountability mechanisms.⁵ **PEER-REVIEWED** His conclusion was that principles alone cannot guarantee ethical AI.

Approval committees can be read as an organisation's attempt to supply two of those missing pieces locally, a method and a point of accountability. That reading is mine. It explains why committees matter, and also why a committee that meets once supplies accountability for a moment and a method that stops at the door.

Morley and colleagues took up the question of method directly. Their review of publicly available AI ethics tools and methods mapped them against ethical principles and against the stages of machine learning development, and found the coverage uneven, with some combinations of principle and stage well served and others barely served at all.⁶ **PEER-REVIEWED** They framed the task as moving from what to how. For an approval body the pressing question is whether any such tools reach the stages after deployment, where most of the change described in the previous chapter happens, and the review offers little reassurance that the supply there was plentiful.

3.2 Audit as a process

Auditing offered a way to turn method into routine. Raji and colleagues proposed an end-to-end framework for internal algorithmic auditing, organised in five stages of scoping, mapping, artefact collection, testing and reflection, and designed to run alongside development so that the audit record is built before a system is deployed.⁷ **PEER-REVIEWED** Among the artefacts it draws on are model cards, the short documents proposed a year earlier by Mitchell and colleagues to accompany trained models with their intended use and their performance across relevant groups.¹⁵ **PEER-REVIEWED** Both ideas are valuable to an approval body. Both, as designed, describe the system at a point in time.

Mökander and Floridi argued, in a short commentary, that ethics-based auditing could help close the distance between principles and practice, treating it as a structured process in which an entity's behaviour is assessed for consistency with relevant

principles or norms, and argued that, to be feasible and effective, such auditing should be a continuous process.¹⁶ **PEER-REVIEWED** The emphasis on behaviour is the important one. An audit on this definition looks at conduct, and conduct continues after approval.

Minkkinen, Laine and Mäntymäki went a step further and conceptualised continuous auditing of AI, adapting an idea long familiar in financial auditing in which assurance is produced on an ongoing basis, close to real time and with a high degree of automation, and they assessed a set of existing tools and frameworks against that concept.⁸ **PEER-REVIEWED** The concept is well formed. What I could not find, when this paper closed, was published evidence on how continuous auditing performs inside ordinary organisations with modest technical teams, which is where most approval committees sit. **EVIDENCE GAP**

3.3 Impacts are made during assessment

Metcalf and colleagues offer the most searching account of why a single assessment is fragile. Studying algorithmic impact assessment alongside older impact assessment regimes, they argue that impacts are evaluative constructs standing in for harms, and that those constructs are co-constructed by the parties involved in the assessment.⁹ **PEER-REVIEWED** Accountability, on their account, depends on there being a forum to which the assessor must answer.⁹ **PEER-REVIEWED**

The consequence for approval, as I read it, is that an assessment carried out once captures the construction of harm that was possible at that date, with the people who were in the room and the uses that were then in view. When a tool's use widens, so does the set of people who might be harmed. The construct in the approval file does not move. Continuous assurance is partly a way of keeping the forum open, so that the construction of impacts can be revisited when the configuration changes.

CHAPTER FOUR

What the frameworks and regulators expect

The language of the lifecycle is already in the main frameworks and in UK policy. What none of them yet supplies is a rule for when an approved system must come back.

4.1 The lifecycle in the frameworks

The National Institute of Standards and Technology (NIST) published its AI Risk Management Framework on 26 January 2023. The framework is voluntary. It organises risk management into four functions, govern, map, measure and manage, with governance running across the other three.¹ **STANDARD** The framework states that risk management should be continuous, timely and performed throughout the dimensions of the AI system lifecycle.¹ **STANDARD**

Several of its subcategories read almost as a specification for what an approval should require after sign-off. One calls for post-deployment monitoring plans that include ways of capturing input from users, appeal and override, decommissioning, incident response, recovery and change management; another expects incidents and errors to be communicated to relevant parties, including affected communities; a third asks for processes to decommission and phase out AI systems safely.¹ **STANDARD** The framework also expects risks from third-party resources to be monitored regularly and pre-trained models to be monitored as part of routine maintenance, which speaks directly to the hosted model problem discussed in the next chapter.¹ **STANDARD**

ISO/IEC 23894, published in February 2023, gives guidance to organisations that develop, produce, deploy or use AI systems on managing the risks specific to AI, building on the general risk management guidance of ISO 31000 and relating risk management activity to the stages of an AI system's life.¹⁰ **STANDARD** Neither document, being a general framework, tells a committee which changes to a particular system should bring it back for a fresh decision, and that gap is where most of the practical difficulty lies.

4.2 The UK position in spring 2023

The government's white paper, published on 29 March 2023, proposes five cross-sectoral principles for existing regulators to interpret and apply within their remits, without placing them on a statutory footing at first, and it anticipates that a statutory duty on regulators to have due regard to the principles may follow.¹¹ **POLICY** Its safety, security and robustness principle expects AI systems to function safely throughout the AI lifecycle, with risks continually identified, assessed and managed.¹¹ **POLICY** The white paper thus endorses lifecycle management while leaving its operational form to regulators and the organisations they oversee.

Data protection already contains a version of the trigger idea. Article 35(11) of the General Data Protection Regulation, as retained in UK law, provides that where necessary the controller shall review whether processing is carried out in accordance with the data protection impact assessment, at least when there is a change in the risk the processing represents.¹⁷ **POLICY** My own view, which goes beyond what the provision says, is that most of the difficulty with adaptive AI lies in recognising when that change of risk has occurred, because the changes described in Chapter 2 rarely announce themselves as changes in risk.

Two other developments shaped the context. On 4 May 2023 the Competition and Markets Authority (CMA) opened an initial review of AI foundation models, examining competition and consumer protection considerations in their development and use.¹⁸ **POLICY** On 22 May 2023 OpenAI's leaders published a proposal for international coordination and oversight of the most capable systems, while arguing that models below a significant capability threshold should be left free of burdensome measures such as licences or audits.¹⁹ **INDUSTRY** I read the CMA review as a sign of official attention to the upstream suppliers from whom many deployers' model changes originate. The OpenAI proposal, whatever its merits, concentrates attention on the most capable systems, while the systems most organisations approve are ordinary ones whose changes are mundane and frequent.

FOR COUNSEL

Whether a supplier's model update, or an informal new use, amounts to a change in the risk represented by processing is a question for the organisation's data protection adviser. The paper does not answer it.

4.3 Change control from medical devices

The most developed answer to the change problem comes from medical devices. In April 2019 the FDA published a discussion paper proposing a lifecycle approach to modifications of AI and machine learning software used as a medical device, under which a manufacturer could set out in advance the modifications it anticipated, which the paper called SaMD pre-specifications, together with an algorithm change protocol describing how those modifications would be made and validated.¹² **POLICY**

The FDA's notice of 3 April 2023 announced draft guidance developing that proposal into a predetermined change control plan. According to the notice, such a plan describes the anticipated modifications and the methodology for implementing them, and is reviewed as part of the marketing submission so that each described modification does not require a further submission; the notice also records that Congress had given the agency express authority to approve or clear such plans in December 2022, and states that the draft is not for implementation.¹³ **POLICY**

What transfers from this, in my view, is the structure. The regulator approves the system together with an envelope of expected change and a method for checking changes inside it, and anything outside the envelope takes the ordinary route. **ECAVEO** An approval committee in a housing association or a university has no statutory powers of that kind, but it can adopt the same structure as a condition of approval.

Instrument	When the assessment happens	What brings the system back
Data protection impact assessment	Before processing begins	Review where necessary, at least on a change in risk

Instrument	When the assessment happens	What brings the system back
Internal algorithmic audit (SMACTR)	During development, before deployment	Not specified by the framework
NIST AI RMF 1.0	Throughout the lifecycle	Post-deployment monitoring and change management plans
FDA predetermined change control plan (draft)	At marketing submission, covering future change	Modifications not described in the plan
Continuous auditing of AI	On an ongoing basis	Built into routine operation

TABLE 1 How five assessment instruments treat change after the initial decision. Source: author's summary of GDPR Article 35(11), Raji et al. (2020), NIST (2023), FDA (2023) and Minkinen, Laine and Mäntymäki (2022).

A futuristic illustration of the Earth, centered in the frame. The Earth is depicted with realistic blue oceans and brown landmasses. Surrounding the Earth are several glowing, translucent orbital rings in shades of blue and orange. Floating around these rings are several small, glowing spheres, some of which are also surrounded by their own smaller rings. Three rectangular, translucent screens are positioned around the Earth, each displaying a line graph with multiple data series. The background is a deep, dark blue with a subtle, swirling pattern, suggesting a cosmic or digital environment.

PART THREE

A working model

A proportionate model of continuous assurance, and what it asks of boards, committees and research ethics panels.

CHAPTER FIVE

A continuous assurance model

Continuous assurance means a small set of commitments written into the approval itself, each with an owner. It does not mean continuous paperwork.

5.1 Nine elements

The model I propose has nine elements. None is new on its own, and several appear in the frameworks already discussed; the proposal is to make them conditions of approval, so that the approval carries its own means of staying current. **ECAVEO**

- A versioned system description, in the spirit of a model card, recording the model and its version, the data it reads, its intended use and the groups it affects.
- A change log that captures all four routes of change, including edits to prompts and settings.
- A short list of reassessment triggers, agreed at the moment of approval.
- Performance monitoring against measures defined at approval and reported on a fixed schedule.
- Bias monitoring across the groups named in the system description.
- A route for reporting incidents and near-misses that is open to staff and to the people affected.
- Renewed consent, or at least renewed notice, when the purpose of use changes.
- A named owner for supplier changes, who reads the supplier's notices and decides whether a trigger has been met.
- Retirement and rollback criteria, set before they are needed.

The versioned description is the anchor for the rest. Mitchell and colleagues designed model cards to travel with a trained model,¹⁵ **PEER-REVIEWED** and an approval body can ask for something similar at the level of the deployed system, with a version number that changes whenever the log records a change of substance. **ECAVEO** The committee then approves a numbered version, which makes it possible to ask, a year later, which version is running and how far it has moved.

1	Describe	Record the approved configuration as a numbered system description.
2	Log	Capture every change, from supplier updates to prompt edits.
3	Test against triggers	Compare each logged change with the triggers agreed at approval.
4	Monitor	Report performance and bias measures on a fixed schedule.

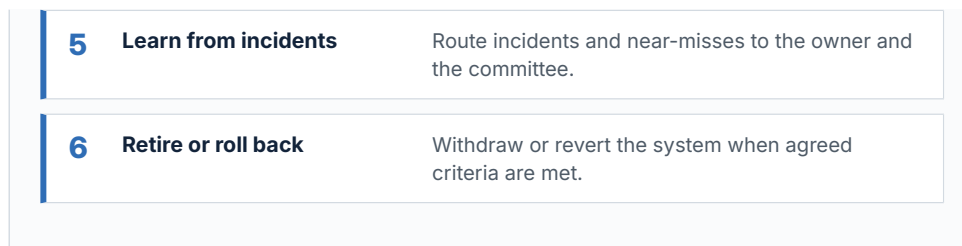


FIGURE 2 The continuous assurance cycle that follows approval. Source: author's framework.

Most of the monitoring elements correspond closely to what the NIST framework already asks of post-deployment plans,¹ **STANDARD** and to the monitoring tests in the rubric proposed by Breck and colleagues.¹⁴ **PEER-REVIEWED** What the model adds is a tie to a specific approval and to a committee that will see the results.

5.2 Triggers for re-review

Triggers are the element that does the most work, and the one committees find hardest to write. The draft list I use with approval bodies is set out below, and I would expect each committee to adapt it. **ECAVEO**

- The underlying model is replaced, or its supplier announces a new major version.
- A new data source, field or feed is added to what the system reads.
- The system is used for a new purpose, by a new team, or on a new group of people.
- A monitoring measure crosses a threshold agreed at approval.
- A serious incident occurs, or near-misses of the same kind accumulate.
- The supplier changes its terms on data use, retention or location.

A trigger is a promise, made at approval, about when the committee will look again. Writing triggers into the approval makes review checkable. A change that meets a trigger comes back for a decision; one that does not is logged and reported in the next summary. In my experience a committee that works this way sees fewer papers, and the papers it does see concern changes that matter.

The obvious objection is that a list like this leaves the hardest judgement, whether a given change is material, to the person who owns the log. That is true, and I do not think any list removes it. The FDA's approach narrows the judgement by requiring the envelope of change to be described in advance,¹³ **POLICY** and the data protection provision frames it as a change in risk.¹⁷ **POLICY** Neither offers a test that a committee could apply mechanically to a hosted language model whose supplier has issued a new version, and I found no published evidence, as at the end of May 2023, on which kinds of change to such systems most often alter their behaviour in ways that matter to affected people. **EVIDENCE GAP**

5.3 Suppliers and the hosted model

Many organisations now approve tools whose models they neither train nor host. For them the first route to change sits almost entirely outside their control, and the NIST framework's expectation that third-party risks and pre-trained models be monitored regularly describes a duty that is easier to state than to discharge.¹ **STANDARD** Supplier notices, where they exist, are usually written for developers, and they rarely say what a new version does differently for a particular use. **PRACTICE**

The practical response has two parts. The first is to name, inside the deploying organisation, the person who reads those notices and decides whether a trigger has been met, since an unowned inbox is how the scenario's supplier upgrade went unnoticed. The second is to ask suppliers, at procurement and at renewal, for a change log at the level the deployer needs, with notice of new versions and some ability to stay on a known version for a period. **ECAVEO** Whether suppliers will agree to that, and whether smaller deployers have the bargaining power to ask, is an open question. The CMA's review suggests that the structure of the upstream market is at least on the regulator's agenda.¹⁸ **POLICY**

CHAPTER SIX

Proportionality and research ethics

Assurance that costs more than the risk it manages will be abandoned. The model therefore needs tiers, and research ethics panels need their own version of it.

6.1 Matching effort to impact

The white paper proposes regulating according to the outcomes AI is likely to generate in particular applications, and it presents proportionality as part of its approach.¹¹

POLICY The NIST framework likewise leaves organisations to set their own risk tolerance and to prioritise their effort accordingly.¹ **STANDARD** Even the OpenAI proposal, which is concerned with systems at the opposite end of the scale from most organisational tools, rests on the idea of a threshold below which lighter treatment applies.¹⁹

INDUSTRY A model that ignored that logic would fail on cost first.

I therefore propose four tiers, set at approval according to what the system does to people and reviewed whenever a trigger is met. **ECAVEO**

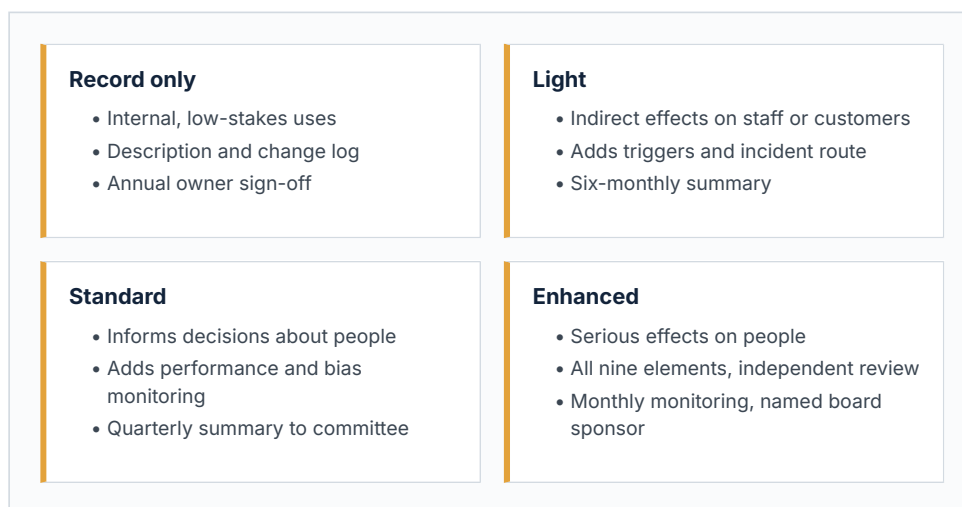


FIGURE 3 Four assurance tiers, set at approval by the system's effect on people. Source: author's framework.

A small organisation can run the lower two tiers with a shared document and a named person, and the change log need be nothing grander than a dated list. What moves a system up a tier is the effect on people, which is why the new-use trigger matters so much in practice (a tool that drafts internal meeting notes and the same tool drafting letters to tenants are different systems for this purpose, whatever the supplier's invoice says). **ECAVEO**

6.2 What research ethics panels might require

Research ethics review is, among all the instruments discussed here, the most firmly tied to a single moment. A panel approves a protocol before recruitment, and changes come back as amendments that researchers are expected to identify and submit.

PRACTICE That arrangement works when the researcher controls the instruments. It works less well when a study relies on a commercial AI tool whose supplier may change it in the middle of data collection.

ILLUSTRATIVE SCENARIO

A doctoral study and a moving tool

An invented doctoral student gains ethics approval for a study in which participants use a commercial chatbot to practise difficult workplace conversations, with transcripts analysed afterwards. The protocol names the tool and describes its behaviour from a pilot. Halfway through recruitment the supplier releases a new version, and the tool begins to respond more cautiously to some scenarios. The student notices the change but does not treat it as an amendment. The panel never hears of it, and the thesis compares two groups of participants who, without anyone intending it, used different systems.

The case raises two problems at once. Participants in the second half of the study consented to use a tool that no longer behaves as the information sheet described, and the study's comparison is weakened by a change nobody recorded. My proposal is that panels considering studies which use adaptive AI tools should require the following as conditions of approval. **ECAVEO**

- The tool is named with its version, and the dates of use are recorded.
- The application states whether the version can be fixed for the study, and what happens if it cannot.
- The application defines which tool changes will be treated as amendments.
- The end-of-study report lists every version change and any effect observed.

These conditions add little to an application, and they move the question of whether a tool update is an amendment from the moment of discovery, when the researcher has every reason to decide it is not, to the moment of approval, when the panel can decide it calmly.

6.3 Consent when purposes change

The seventh element of the model, renewed consent or notice when purposes change, applies both to research and to organisational tools. People who agreed to one use of their information, or of a tool that affects them, agreed to a particular bargain, and a new use changes the bargain even when the software does not change at all. **ECAVEO** In research the route is familiar, through re-consent or a revised information sheet. In organisational settings it usually means a revised notice and, where the effect on people is significant, an honest assessment of whether the original basis for using their information still holds, which is a matter for the organisation's data protection adviser.

CHAPTER SEVEN

Recommendations and open questions

Approval should become the first entry in a continuing record. The recommendations below set out how a board and its committees can make that happen without building a new function.

7.1 What I recommend

The recommendations follow from the argument of the preceding chapters and constitute my own framework. They are addressed to boards and to the committees that act for them. **ECAVEO**

1. Treat approval of any AI system above the lowest tier as approval of a numbered version, with a system description that the named owner keeps current.
2. Write reassessment triggers into every approval, adapted from the list in Chapter 5, and require the owner to report each logged change against them.
3. Set the assurance tier at approval according to the system's effect on people, and move the system between tiers when a trigger is met.
4. Name an owner for supplier changes in every deployment that relies on a hosted model, and ask suppliers at procurement for version notices and a period of version stability.
5. Require scheduled performance and bias reports for systems in the upper two tiers, with measures and thresholds agreed before deployment.
6. Open an incident and near-miss route to the people a system affects as well as to staff, and route its reports to the committee that gave the approval.
7. Agree retirement and rollback criteria at approval, while nobody yet has an interest in keeping the system running.
8. Ask research ethics panels to adopt the four conditions in Chapter 6 for studies that rely on adaptive AI tools.

None of these depends on new legislation. Each draws on expectations already present in the NIST framework,¹ **STANDARD** in the white paper's lifecycle principle,¹¹ **POLICY** and in the change control logic the FDA has been developing for medical devices.¹² **POLICY**

7.2 Conclusion and open questions

A one-off approval of an adaptive AI system describes a configuration that begins to dissolve as soon as the system is used. The research literature has moved steadily towards treating assessment as continuing work, and the main frameworks and UK policy now speak the language of the lifecycle, but the operational bridge between an approval and a system that keeps changing is still largely missing outside medical devices. The model in this paper is one attempt at that bridge, built from commit-

ments light enough for a small organisation to keep. It is untested, and I would expect some of its triggers to prove too broad and others too narrow once committees use them.

Several questions remain open, and the answers will come from practice more than from further argument.

- Which changes to a hosted model should count as material, given that deployers rarely see what a new version does differently?
- Can suppliers be required by contract to provide change logs at the level deployers need, and will smaller deployers have the bargaining power to ask?
- How should a research ethics panel handle a study whose AI tool updates in the middle of data collection when the version cannot be fixed?
- What evidence would show that continuous assurance reduces harm in ordinary organisations, and who is placed to collect it?
- If the model dementia reported by Shumailov and colleagues is confirmed,⁴ **PREPRINT** should bias monitoring give particular weight to the rare cases that successive model versions may lose?

The last question may matter most for the people least able to raise it, since they are the ones who sit in the tails.

END MATTER

References

1. National Institute of Standards and Technology (2023) *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. 26 January 2023.
2. Sculley, D. et al. (2015) Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems* 28.
3. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A. (2014) A survey on concept drift adaptation. *ACM Computing Surveys* 46(4), 1-37.
4. Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N. and Anderson, R. (2023) Model dementia: generated data makes models forget. arXiv:2305.17493v1, 27 May 2023.
5. Mittelstadt, B. (2019) Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1(11), 501-507.
6. Morley, J., Floridi, L., Kinsey, L. and Elhalal, A. (2020) From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics* 26(4), 2141-2168.
7. Raji, I.D. et al. (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33-44.
8. Minkinen, M., Laine, J. and Mäntymäki, M. (2022) Continuous auditing of artificial intelligence: a conceptualization and assessment of tools and frameworks. *Digital Society* 1(3), 21.
9. Metcalf, J., Moss, E., Watkins, E.A., Singh, R. and Elish, M.C. (2021) Algorithmic impact assessments and accountability: the co-construction of impacts. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 735-746.
10. ISO/IEC (2023) *ISO/IEC 23894:2023 Information technology: Artificial intelligence: Guidance on risk management*. Geneva, International Organization for Standardization.
11. Department for Science, Innovation and Technology (2023) *A pro-innovation approach to AI regulation*. CP 815, 29 March 2023.
12. US Food and Drug Administration (2019) *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)*. Discussion paper and request for feedback, April 2019.
13. US Food and Drug Administration (2023) *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions*. Draft guidance, notice of availability, *Federal Register* 88, 19648, 3 April 2023.
14. Breck, E., Cai, S., Nielsen, E., Salib, M. and Sculley, D. (2017) The ML test score: a rubric for ML production readiness and technical debt reduction. *2017 IEEE International Conference on Big Data*, 1123-1132.
15. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T. (2019) Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220-229.
16. Mökander, J. and Floridi, L. (2021) Ethics-based auditing to develop trustworthy AI. *Minds and Machines* 31(2), 323-327.
17. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation), Article 35(11). *Official Journal of the European Union* L 119, 4 May 2016. Retained in UK law as the UK GDPR.
18. Competition and Markets Authority (2023) *AI Foundation Models: Initial Review*. Review launched 4 May 2023.
19. OpenAI (2023) Governance of superintelligence. 22 May 2023.

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

Approved Once, Changed Daily

Ecaveo Working Papers, AI Adoption: The Higher-Order Issues
Paper five of six · 31 May 2023