Monitoring AI models after launch

For chief risk officers, model owners and product committees at insurers, platforms, wealth managers and advice firms: why AI models drift after launch, and how to monitor them the way an actuary monitors experience against a pricing basis.

IN BRIEF

The problem

AI models are tested hard before launch and then left to run, while the customers, products and data they see keep changing.

Why it happens

Budgets and attention go to the build, so nobody agrees in advance what the model should achieve, what would trigger a review, or who would act.

What to do

Before launch, agree a few measures with expected levels and triggers, split them by customer group, and report actual against expected to a committee each quarter.

01

Why a model that passed its tests can stop working

An AI model is approved on the results of its testing: how often it got the answer right on a set of past cases. From the day it goes live, the cases it sees start to differ from the ones it was tested on.

Customers change, products change, a new channel brings a different kind of request, and the wider economy moves. The supplier may update the model underneath the tool. And the people using it change too: as trust grows, staff check its output less closely.

Most of this builds gradually. A model that routes customer requests, flags customers who may need extra support or helps price a risk can drift for a year before anyone notices, and the first sign is often a rise in complaints.

02

Monitor it the way an actuary monitors experience

Insurers have long done this for their pricing. An actuary sets a basis, the assumptions the price relies on, then compares actual experience with expected each quarter or year. The ratio of actual to expected shows whether the assumptions still hold, and a ratio outside an agreed range prompts a review.

An AI model needs the same discipline. Before launch, agree four things.

What the model should achieve. A few measures, each with an expected level taken from testing: accuracy on a sample checked by staff, the share of cases sent to exceptions, how often reviewers overrule the model, and complaints linked to it.

What would trigger a review. A level for each measure, set before the results come in, and a rule for trends that prompts a question before the trigger is reached, such as three quarters in a row above expected.

Which groups to look at separately. A model can hold its overall accuracy while getting worse for one group of customers.

Who owns it and who acts. A named owner, a committee that sees the results each quarter, and a recorded decision on every breach.

> FROM EXPERIENCE
>
> At Aviva I monitored annuity experience and profitability to propose rate changes, and researched socio-economic rating factors for annuity pricing. At Partnership I ran the first equity release experience analysis.
>
> That work taught me to agree what I expected before the results arrived, so I could see when reality had moved, and to look below the total. A book of business can match its assumptions on average while one group of customers moves away from them, and the same is true of a model.

03

An illustrative first year

ILLUSTRATIVE EXAMPLE

One AI model's first year against its launch testing

Quarter	All customers	Customers aged 75 and over	The committee's response
Q1	97%	94%	Noted
Q2	104%	117%	Asked why the older group is rising
Q3	107%	139%	Review, with extra checks meanwhile
Q4	101%	106%	Model retrained; back inside the trigger

Illustrative: a model routing 30,000 customer requests a quarter to the right team, a fifth of them from customers aged 75 and over. Each quarter staff check 2,400 requests from customers under 75 (10%) and 1,800 from customers aged 75 and over (30%); the all-customer figure is weighted back to the full book. Launch testing found 3% routed wrongly in both groups. Figures are errors found as a percentage of those expected; the trigger for review is 125%.

The overall ratio stays inside its trigger all year. The ratio for older customers rises in the second quarter and breaches in the third. Because they make up a fifth of the requests, their error rate would have had to more than double before the total

crossed the trigger. A firm watching only the total would probably have found out from complaints.

Set triggers with the sample size in mind. That is why the older group is checked at three times the rate: at the same 10% as everyone else, its 18 expected errors a quarter would breach a 125% trigger by chance about one quarter in seven.

04 When a trigger is breached

Decide the options before launch, so that a breach leads to a decision rather than a debate: extra human checks while the cause is found, retraining or recalibrating the model, narrowing what it is used for, or switching it off and returning to the previous process. The last option works only if the previous process still exists and someone can still run it.

Some reviews shouldn't wait for a trigger. When a supplier changes the model underneath a tool, or the firm starts using it for a new product or a new group of customers, re-run the launch tests on the firm's own cases before relying on it again.

Two costs belong in the business case from the start: monitoring, which runs for as long as the model is in use, and keeping a fallback in working order.

05 What regulators expect

For banks and building societies with approval to use internal models for regulatory capital, the PRA's model risk management principles (supervisory statement SS1/23, effective since 17 May 2024) expect performance to be monitored against thresholds set when the model was developed, and changes in products, clients or market conditions to be assessed for recalibration. The PRA has said the principles deliberately included factors relevant to AI models. Insurers are outside their scope. The PRA has highlighted AI adoption in its 2026 supervisory priorities, and in April 2026 said it intended to build further on the principles this year, without saying it would extend them to insurers.

For FCA-regulated firms, the Consumer Duty requires firms to monitor outcomes, identify whether any group of customers is getting worse outcomes than another, and act on what they find (the FCA's Principles for Businesses, PRIN 2A.9, as at October 2026).

Questions for the board

01 Which of our AI models are live, who owns each one, and when did a committee last see how it is performing?

02 For each model, what did testing lead us to expect, and how does actual performance compare?

03 Are the results split by customer group, including customers who may be in vulnerable circumstances?

04 What level would trigger a review, and was it agreed before the results came in?

05 If we switched a model off tomorrow, could the previous process take over, and who would run it?

06 How would we know if a supplier changed the model underneath one of our tools?

Launch testing shows how a model performed on yesterday's cases. Monitoring it the way an actuary monitors experience, against expectations agreed in advance and below the level of the total, gives a firm the best chance of seeing drift before its customers feel it.

If this is on your table, a 30-minute conversation is usually enough to tell whether I can help.

Talk to Mark mark@markwoodruff.co.uk · 07391 820971 · markwoodruff.co.uk

ABOUT THE AUTHOR

Mark Woodruff is a Chartered Actuary (FIA) and Chartered Financial Planner (FPFS). He spent 15 years pricing and building products at Aviva, Prudential, Partnership and St. James's Place, then more than seven years founding and running advice firms through to a substantial sale. Through Mark Woodruff Consultancy he works with financial services firms on product, pricing, distribution and governance.

Regulatory notice. Mark Woodruff Consultancy works with businesses only. I don't give regulated personal financial advice to individuals.

Disclaimer. This briefing is general information for financial services firms. It is not tailored to any firm's circumstances and is not legal, regulatory or financial advice. Work is agreed in a written letter of engagement before it starts.

Published October 2026
Regulatory references correct at October 2026
© Mark Woodruff Consultancy 2026